Trends in Undergraduate WxChallenge Performance

NICHOLAS J. ECKSTEIN,[a] NATHAN M. HITCHENS,[b] AND PETRA A. ZIMMERMANN[b]

[a] *Department of Earth and Space Sciences, Pierce College, Lakewood, Washington*
[b] *Department of Geography and Meteorology, Ball State University, Muncie, Indiana*

(Manuscript received 20 September 2022, in final form 25 April 2023, accepted 28 April 2023)

ABSTRACT: The WxChallenge is a national forecasting competition in which participants at colleges and universities across North America attempt to accurately predict daily high and low temperatures, maximum wind speed, and precipitation accumulation at a variety of locations each year. Undergraduate students make up the majority of participants. In this study, we observed trends from 11 seasons of WxChallenge data and related them to existing literature on local forecasting contests. Normalized scores were calculated each day for any participant that submitted a forecast. On average, undergraduate scores improved with continued participation in the contest. Significant gains are made during their first year by forecasters who participated for multiple years. Duration of participation in the contest plays a role, but significant improvements in performance were also noted with higher academic standing, potentially as a result of forecasting experience gained through other curricular or extracurricular activities.

SIGNIFICANCE STATEMENT: One avenue that aspiring meteorologists have outside the classroom to hone their forecasting skills is through the WxChallenge, a national collegiate forecasting contest. We wanted to see whether, on average, students who participated longer in the contest outperformed those who did not forecast for as long. We found this to be true. Mean scores by students who participated for more than one year improved significantly by the end of the first year, and they outperformed those without prior experience in the WxChallenge in following years. These findings are important because they can be used as evidence to motivate students to join a forecasting contest early in their academic careers, allowing them to gain skill before pursuing a career in forecasting.

KEYWORDS: Forecast verification/skill; Forecasting; Education

1. Introduction

With weather affecting a significant portion of the gross domestic product of the United States (AMS Council 2007), serving as the focus of major stories on local media (Lorditch 2009; Wilson 2008), as well as being responsible for the majority of emergencies (AMS Council 2015), it behooves a society to have well-trained meteorologists with solid weather prediction skills. Certainly, the chaotic nature of the atmosphere makes weather forecasting a challenging endeavor. As meteorological technology and statistical techniques evolve, weather models [numerical weather prediction (NWP)] advance to achieve greater accuracy; however, human forecasters have frequently improved upon model output, and the human element remains essential to the forecasting process (AMS Council 2015). Meteorology and atmospheric science educators, tasked with training future meteorologists, seek approaches to integrate NWP and students' understanding of the behavior of the atmosphere from their standard coursework. One activity, the weather forecasting contest, immerses students into environments with elements similar to professional ones and may serve as a vital training complement to content-based education. Such forecast contests are not unique and have been conducted locally at numerous colleges and universities (e.g., Hilliker 2008; Roebber and Bosart 1996; Suess et al. 2013), and even nationally (Peyrefitte and Mogil 1968) for decades. While the impact of local contests has been studied many times over the years, the effect that national contests have on student forecasting performance has not yet been analyzed. In this study undergraduate student scores from the current national weather forecasting contest, WxChallenge, were examined to assess trends that could be used to inform how we train future forecasters.

With the successes of NWP, forecasters increasingly rely on its output. Consequently, roles of forecasters have been transformed accordingly. Stuart et al. (2006) summarized the continuing evolution of these roles and suggested that operational forecasters may lack comprehension of some aspects of NWP models due to insufficient time for learning. They noted the importance of continued education with a foundation of critical thinking. Objective NWP output does reveal atmospheric phenomena imperceptible to even the most skillful forecasters, with model output often regarded as "truth" by many forecasters (Boi and Spangler 2012). Still, Stuart et al. (2006) suggested that all operational forecasters lack a solid understanding of at least some aspects of NWP models and that increased education about NWP is necessary for university students to gain a more solid basis on strengths and weaknesses of the models.

Despite the clear benefits and advantages of weather models, private forecasting firms have generally felt that graduating meteorologists depend too much on these model forecasts; students should experience forecasting without the assistance of this guidance to learn meteorological principles (Yarger et al. 2000). To address this, some universities offer forecasting classes, laboratories, and other ancillary activities to teach

Corresponding author: Nathan Hitchens, nmhitchens@bsu.edu

DOI: 10.1175/WCAS-D-22-0102.1

forecasting processes. These include field activities at the U.S. Naval Academy (Barrett and Woods 2012), high-altitude balloon launches at Ball State University (Coleman and Mitchell 2014), a student research study at Texas A&M on hurricane perception (Morss and Zhang 2008), and The Pennsylvania State University's online certificate program in weather forecasting (Grenci et al. 2008).

Meteorology students remain captivated by weather prediction. Knox and Ackerman (2005) surveyed students at the University of Georgia and the University of Wisconsin and found that weather forecasting generated the most interest, surpassing even severe weather (e.g., tornadoes). This invites serious concern, because textbooks often relegate forecasting to a single chapter at the end and the included information may not be up to date (Knox and Ackerman 2005).

Active learning engages student learners more directly than traditional lecture, requiring them to participate in meaningful activities and think about said activities (Bonwell and Eison 1991). Going beyond mere memorization and recall of facts, many active learning strategies necessitate student employment of higher-level thinking skills (e.g., analysis, evaluation) according to Bloom's taxonomy (Bloom 1956). Freeman et al. (2014) found improved examination scores for students in science, technology, engineering, and mathematics classes with active learning components (as compared with traditional lecture); these findings suggest that potential modifications in the pedagogy of sciences, including meteorology, focus more on activity-based learning strategies. Participation in forecasting contests, in which students synthesize their knowledge in the context of problem-solving (i.e., predicted weather in a given location) serves as one active learning approach in meteorological education, engaging them in a manner that mimics their eventual professional employment.

Whether during or external to class time, a weather forecasting contest affords students a creative opportunity to solve that most prosaic of questions: what will the weather be? Numerous contests have been held, from local to national levels, with participants ranging from those enrolled in a single meteorology class to thousands of students from nearly 100 universities across the United States (Illston et al. 2013). Local contests (i.e., those constrained for use in a single university) serve to engage students actively and have shown some interesting results. In one of the earliest studies on student performance in a local forecasting contest, Gedzelman (1978) examined the results of the City College of the City University of New York's forecast game, finding that those participants with less experience (one prior meteorology class taken) had similar scores as those with more experience by the 30th forecast, although those with more experience still performed better in unusual weather situations. The sample size for this study was small (12 participants), so later work such as that of Hilliker (2008) and Suess et al. (2013) helps to further explore the topic of improvement by beginning forecasters. Nonmajor students in an introductory meteorology course improved their skill in a forecasting contest at Iowa State University over the first 10–15 forecasts, but then little additional improvement was observed thereafter (Suess et al. 2013). These students performed slightly better than a persistence forecast. Relative to weather model forecasts, Hilliker (2008) found that student performance (both meteorology majors and nonmajors) increased throughout the semester (9 forecasts).

The issue of forecasting experience, such as that examined in Gedzelman (1978), was further explored by Roebber and Bosart (1996). They subjectively labeled each forecaster as either high or low experience based on their familiarity with forecasting products, general forecasting experience, and overall curiosity about the weather; these criteria were separate from the amount of formal education each forecaster had. Interestingly, they found that differences in skill were significant based on the experience level of the forecasters, whereas no such significance was observed when comparing education. As with Gedzelman (1978), these authors noted an increase in skill by inexperienced forecasters over time, with a main advantage enjoyed by experienced forecasters of identifying situations in which usual forecasting rules do not apply, and adjusting accordingly.

At Iowa State University, introductory meteorology students were required to forecast specific weather parameters and provide physical reasoning for their forecasts (Yarger et al. 2000). The authors of the study found that including these explanations engaged students, permitted understanding of factors affecting weather elements, and required students to determine causes for future weather beyond values from a website or model; the students indicated that this was the most popular component of the introductory meteorology course, and they believed it greatly improved their skills (Yarger et al. 2000). Similarly, at West Chester University, a forecasting contest for introductory classes was designed to bridge the gap between the theoretical and operational aspects of meteorology, increasing comprehension of weather prediction (Hilliker 2008). It was found that participation decreased with time but increased with course level, with improved accuracy each semester, eventually outdoing even computer model forecasts.

Some contests employ probabilistic forecasting (Decker 2012; Hamill and Wilks 1995). The New Brunswick Forecasting Game (Rutgers, The State University of New Jersey), based on the generation of probabilistic forecasts, provided students with exposure to the communication of uncertainty; students performing well in the local contest also tend to do well in national forecasting contests (Decker 2012). Cornell University's probabilistic forecast contest demonstrated that assessment of daily forecast uncertainty is difficult and that objective methods should be used to quantify forecast uncertainty (Hamill and Wilks 1995).

From its beginnings as a competition between Florida State University and the University of California, Los Angeles (UCLA), the National Collegiate Weather Forecasting Contest (NCWFC), forerunner to the WxChallenge, evolved into its recent form by the early 1980s (Roebber et al. 1996). Post-NCWFC, the WxChallenge began at the University of Oklahoma in 2005, expanding to a full contest in 2006 (Illston et al. 2013). This competition takes advantage of technological advances in consumer electronics and meteorology while introducing students to a variety of forecast locations and issues.

The structure of the WxChallenge permits competition between forecasters of similar educational standing, as well as with the entire group; additionally, forecasters can be registered by course, allowing instructors to compare scores within a given course. To date, nearly 100 universities and over 2000 forecasters have participated in a single year.

Whether success in a local forecast contest translates to academic performance is unclear. Cervato et al. (2011) found that those students in an introductory meteorology course that began forecasting in their local contest early in the semester and continued forecasting throughout were significantly more successful in the course. Likewise, Hilliker (2008) noted a significant negative trend between a student's contest ranking and their course grade—in the case of rankings lower (such as 1) is better—positing that academic ethic may be motivating their success (e.g., more time spent constructing forecasts). However, Schultz et al. (2013) observed that, among the third-year students with little meteorological experience that they studied, those who were better academically did not always perform best in the local forecast contest. Course grades are not the only outcome from participation in a forecast contest, as the authors stated that through this endeavor students were engaged in critical thinking and higher-level skills from Bloom's taxonomy.

Both local and national forecasting contests have been operating for many years and have become a staple in meteorology programs and courses. Studies of student performance in forecasting contests have been focused on local contests; Skeeter (2006) suggested that local contests have advantages over national contests, including a better chance for an individual to win and more engagement by students if forecasting for their local area. Indeed, Hilliker (2008) found that the primary reasons for a lack of student participation in their contest were forgetfulness and discouragement. Since there has been little research focusing on national forecasting contests, in this study we ask, "What trends can we see in 11 years of WxChallenge forecast contest data?" To address this, we analyzed the results of undergraduates participating in the WxChallenge, specifically focusing on changes in performance over time relative to experience. Here we considered experience both relative to the contest itself (i.e., years of prior participation) and within the academic setting (i.e., class standing). Additionally, we examined whether participation as part of a course (likely compulsory) affected performance as compared with those voluntarily participating. Our study utilized data from the archives of the contest itself to answer these questions.

2. Data and methods

Forecast data were downloaded directly from the WxChallenge website and spanned the years 2005–16. Over these 11 forecast contest seasons, each covering the academic fall and spring semesters, forecasters from more than 120 colleges and universities contributed to an average of more than 1700 participants each season. With 8 forecast days per location, and 10 different locations making up the regular season, plus 12–16 additional days for a season-ending tournament location, depending on the season, there were 1040 forecast days available to analyze.

Each forecast contained the following elements: forecaster identifier (ID), academic institution, forecaster category (i.e., class standing), class participation (i.e., if their participation was part of a course the student was taking at their institution), forecast type, high temperature, low temperature, maximum sustained wind speed, and quantitative precipitation forecast. Upon registering for WxChallenge, each forecaster chose a category based upon their current status in the academy: freshman/sophomore (category 4), junior/senior (category 3), graduate student (category 2), faculty/staff/postdoctorate (category 1), or professional (e.g., alumni; category 0). Even if a participant's status changed during the academic year (i.e., between the fall and spring semesters), their category did not change. For this study, only undergraduate forecasters were considered.

The WxChallenge allows for four types of forecasts: 1) numerical forecasts, which consist of values entered by the forecaster; 2) guidance forecasts, which use values from the 1800 UTC run of the Global Forecast System (GFS) weather model; 3) persistence forecasts, which take the previous day's values and use them for the current forecast (this type was discontinued in 2020–21); and 4) missed forecasts, which use the climatological values for the forecast location as the current forecast. Since the latter three forecast types do not reflect any ability by the forecaster to accurately predict the weather, and result in the forecaster being assessed error points in the contest, only numerical forecasts were used in this study.

Each forecast element (i.e., high and low temperature, wind, and precipitation) was predicted by forecasters four days per week (from Monday through Thursday) for 2 weeks at each location. The 24-h period over which a single forecast was valid was from 0600 UTC the day after the forecast was submitted (e.g., Tuesday for a forecast submitted on Monday) through 0600 UTC the following day for temperatures and precipitation and from midnight to midnight local standard time for wind. During this time period, a forecast is intended to accurately predict the maximum and minimum temperatures (°F), the maximum sustained wind speed in knots (kt; 1 kt $\approx$ 0.51 m s^{-1}), and cumulative liquid-equivalent precipitation (to the nearest 0.01 in.; 1 in. = 2.54 cm). Locations were chosen as forecasting sites based upon the existence of a National Weather Service (NWS) weather station there. The measured values of the four forecast elements were used as the basis to evaluate all forecasts, and the values were recorded on the WxChallenge web page.

Although it may seem straightforward, determining the class standing of an undergraduate student participating in the WxChallenge was a nontrivial task. Most students only participated for all or part of one season (about 70% of all undergraduates), so there exists a level of ambiguity inherent in the forecaster category they identified with that season. For instance, a category 4 forecaster could be a freshman or a sophomore. The class standings of students who participated for two or more seasons become easier to define: two consecutive seasons as a category 4 forecaster most likely indicates that student was a freshman their first season, and a sophomore

the next. Similarly, a student who participated in two consecutive seasons, first as a category 4 forecaster then the next year as category 3, suggests the student was first a sophomore, then entered the next season as a junior. The only limitations to following a student through multiple seasons of the WxChallenge are whether they chose the same forecaster ID each time or transferred to a different university (the former being much more likely). Since forecaster ID and academic institution are the only means by which participants can be identified, a change in ID or university would result in wrongly believing more than one student participated for a single season rather than one student over multiple seasons. Unfortunately, these issues surrounding single-season participants—both the lack of specificity of their class standing and the possibility of multiseason participation classified incorrectly due to an ID or university change—must be accepted as possibilities when interpreting results, since there is no way of resolving them. Additionally, class standing is typically defined by the number of credits a student has earned at an institution and may not be representative of a student's meteorological education or experience.

Scoring in the WxChallenge uses a system by which error points are assigned based upon the differences between values forecast by participants for each of the four elements and the values actually measured for each particular forecast day. One error point is assigned for every degree difference for temperature forecasts (both maximum and minimum temperatures), one-half of an error point is assigned for every knot difference for wind speed, and a variable scale of error points (ranging from 0.4 to 0.1 depending on the total and forecast precipitation for that day) is assigned for every difference of 0.01 in. of liquid precipitation. The sum of error points from these four forecast elements composed the raw score for a forecaster for each day, with the overall goal of minimizing the number of error points by forecasting as closely to what will happen as possible.

Since the weather at forecast locations used in the WxChallenge varied in the level of challenge from site to site and week to week, the raw scores for each forecaster for each day were normalized in order to compare between different days and different locations. A normalized score was calculated using the following formula:

$$\text{normalized score} = \frac{(\text{raw score} - \text{consensus score}) \times 10}{\text{standard deviation}}, \tag{1}$$

where the consensus score was the score achieved by averaging the forecasts for each of the four forecast elements using only human-submitted forecasts and assigning error points to these average forecasts based on the scoring system described above. The standard deviation was calculated using each of the raw scores from human-submitted forecasts. Based on this normalization method, individual normalized scores that were better than the consensus score for a day had negative values (i.e., the raw score was lower than the consensus score) and those that were not were positive. This is the normalization method actively used by the WxChallenge to rank forecasters

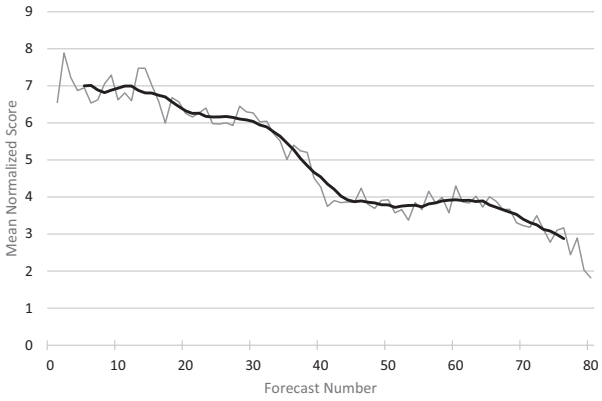

FIG. 1. Mean normalized score by ordinal forecast for single-year forecasters. The gray curve is the value for each ordinal forecast, and the black curve is the centered 9-day mean value.

for each 2-week forecast period, as well as overall during each contest season. Forecast skill is defined as the accuracy of a forecast relative to the accuracy of a forecast produced by some standard of reference (Murphy 1993). Here we chose the consensus forecast as the standard of reference in order to mimic the scores used by students and instructors from the WxChallenge, although it is certainly possible to choose other forecasts as the standard of reference on which to base forecast skill.

When aggregating forecasts for analysis in this study, only human-submitted forecasts were used, as they are the only forecast type that reflects the meteorological knowledge and skills of an individual forecaster. This means that when examining performance chronologically, days on which nonhuman forecasts were submitted by a forecaster were removed, leaving only a time series of human-submitted forecasts with no gaps. Likewise, data aggregated by group (such as class standing or year of participation) also omitted any nonhuman forecasts from members of that group.

3. Results and discussion

Many factors conceivably affect WxChallenge participant skills, including class standing, amount of participation in a single contest year, number of years students participated in the WxChallenge, or whether participation was part of a formal meteorology class. Beginning with participation over a single year, it is seen that these forecasters improved across the duration of the 80 forecasts that make up a season (Fig. 1). This improvement was not uniform, as there appears to be a period of modest improvement through the first four cities (32 forecasts), followed by a relatively large decrease in the mean normalized score by about 2 points over the next approximately 8–12 forecasts. After that the mean score remained relatively constant until approximately forecast 65, followed by another modest decrease. Indeed, through participation during just one season, gains in forecasting skill were realized, so long as the forecasters persisted through the first semester (40 forecasts if none were missed). These results are consistent with Gedzelman (1978), who noted improvement

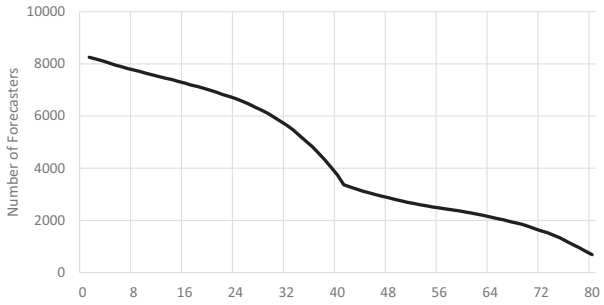

FIG. 2. Number of single-year forecasters as based on the forecast number.

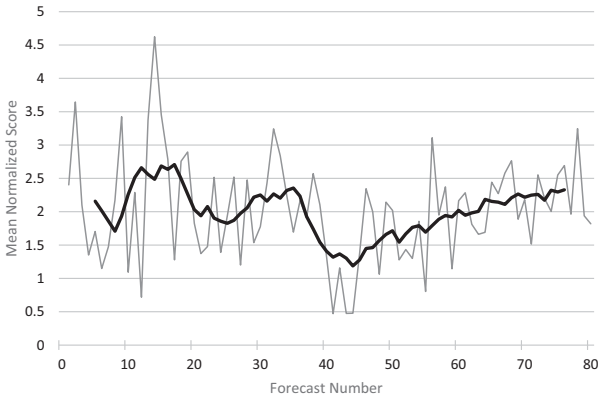

FIG. 3. Mean normalized score by ordinal forecast for participants who submitted 80 human-generated forecasts. The gray curve is the value for each ordinal forecast, and the black curve is the centered 9-day mean value.

over the first 30 forecasts, and Suess et al. (2013), who observed early improvements (albeit over the first 10–15 forecasts), with a plateau in skill following that initial improvement.

A one-way analysis of variance (ANOVA) test between average scores from forecasts 1–9 (beginning of season), 36–44 (middle of season), and 72–80 (end of season) revealed a significant difference at a 1% significance level, and a Tukey test comparing each pairing of these averages (with a 95% experiment-wise confidence level) identified each of the three pairings as significant at the same 1% level. However, an examination of the number of forecasters who submitted a human-generated forecast throughout the season shows a steady decrease through the first semester, accelerating somewhat during the last period (forecasts 33–40), and again in the second semester (Fig. 2). This could explain some of the improvement in scores throughout the year, especially near the end of the first semester.

While specific reasons for the decline in participation and relation to increased mean performance cannot be ascertained from this dataset, aside from forgetfulness, Hilliker (2008) reported that discouragement was a leading reason for a lack of participation in their contest. In this case, there could be a subset of students that struggled to remember to forecast several times, and as a result became discouraged by the resulting poor scores and stopped forecasting. Focusing on discouragement alone, there could also have been some students that tried forecasting, did not do well, and after several times seeing this happen decided to stop forecasting. This latter scenario speaks to the learning theory of self-efficacy (Bandura 1977), which describes a person's belief that they can do something; here it is their belief that they can forecast. Additionally, it has been found that students' motivation is significantly correlated with their attitude toward and achievement in science (Tuan et al. 2005). High motivators have been found to include grade motivation, career motivation, and intrinsic motivation (Chumbley et al. 2022). In this instance, perhaps those students that forecast more often through the year did so because they were meteorology majors and were motivated by career aspirations.

In an attempt to better understand whether single-year participants improved during the course of a WxChallenge season, the analysis was repeated using scores from only the 688 students who submitted a human-generated forecast for all 80 days (Fig. 3). In this case, the mean normalized score improved by the end of the first semester, but then became increasingly worse throughout the second semester. An ANOVA test comparing 9-day averages representing the beginning, middle, and end of the season (as above) was significant at the 1% level, and a Tukey test revealed significant differences between the mean normalized scores from the middle of the season compared to both the beginning and end at the 5% and 1% levels, respectively. The greatest improvements occurred during the third city (forecasts 17–24) and the second half of the last city of the first semester through the first half of the first city of the second semester of the season (forecasts 37–44). Since nonparticipation was controlled for here, we know that is not a factor, but reasons for the trends in the mean normalized score are unknown and can only be speculated. Since the continual decline in performance coincides with the majority of the second semester, one possibility is that these students are putting less effort into their second-semester forecasts, perhaps due to mental fatigue or more challenging classes being taken in the spring semester, leading these students to spend less time on the WxChallenge. An additional explanation may be that with far fewer undergraduate students participating during the second half of the season there was a relatively larger share of faculty, postdocs, and graduate students participating, all of whom were likely to forecast more accurately as a group than undergraduates due to their advanced meteorological knowledge and greater forecasting experience. This is important because the normalized scores are based on the consensus score from all human-submitted forecasts, so by reducing the number of undergraduates, better normalized scores may have been more challenging for any forecaster to achieve, let alone undergraduates who ultimately only participated for one year.

Much like the single-year forecasters, those who participated over multiple years demonstrated improvements in their forecasting abilities over time (Fig. 4), with a relatively large gain near the end of their first semester of participation (by approximately the 40-forecast mark). These forecasters

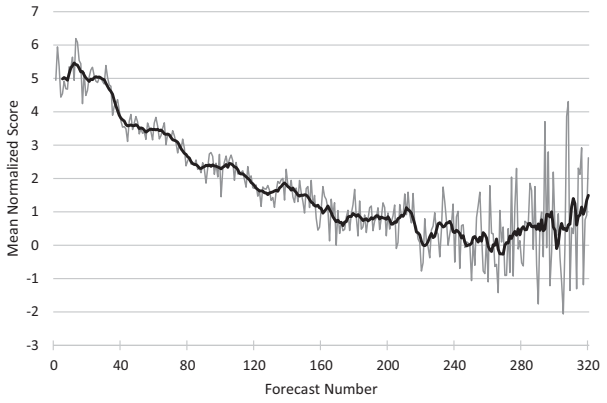

FIG. 4. Mean normalized score by ordinal forecast for multiyear forecasters. The gray curve is the value for each ordinal forecast, and the black curve is the centered 9-day mean value.

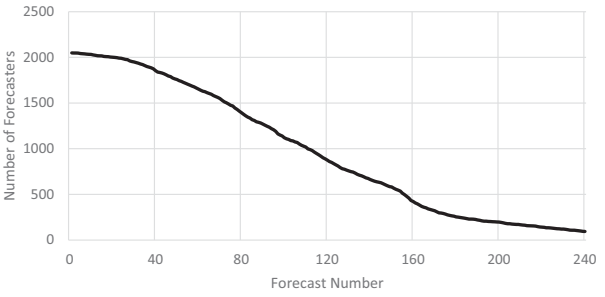

FIG. 5. Number of multiyear forecasters as based on the forecast number.

experienced another substantial improvement in skill during their second semester and continued to see modest gains through their third year of participation (around the 240th forecast). After this point the variability of the mean scores becomes increasingly large, likely because the number of forecasters who participated more than three years, let alone without missing many forecasts, is relatively small (94 forecasters had at least 240 forecasts). A one-way ANOVA test between the average scores of forecasts 1–9, 76–84, 156–164, and 236–244 (approximating the beginning of year 1, the end of year 1, the end of year 2, and the end of year 3) revealed significant differences between these groups at the 1% significance level. A Tukey test between each of the six possible pairs of forecast scores revealed significant differences at the 1% level for all combinations with the exception of those representing the ends of years 2 and 3. These results support the idea that continued forecasting experience leads to continued improvement, at least through the first two years of participation in the WxChallenge. As with single-year forecasters, the number of students submitting forecasts decreased continually over time (Fig. 5), leading to questions about whether these results are influenced by this decline in participation.

In the case of multiyear forecasters, there were 94 students that submitted at least 240 forecasts (the equivalent of three full years of WxChallenge participation). Using this subset to examine performance over time, a pattern similar to that from using all forecasters is revealed, with an apparent improvement over the first year (notably from forecasts 1–40 and 40–80), and gradual improvements over the remaining two years (Fig. 6). An ANOVA test comparing the mean scores at the beginning of year 1, then at the ends of years 1, 2, and 3 (see above) reported a significant difference at the 1% level. A Tukey test identified significant differences between the mean score from the beginning of year 1 to the mean scores from the ends of each of the three years at a 1% level, but the differences between the scores from the ends of the three years were not significant. This suggests that the mean scores from these students improved significantly over the course of their first year, but their mean scores remained nearly the same across the remaining two years. Although the sample is much smaller, these results differ from the subset of single-year students, with a significant improvement in mean scores by the end of their first year. One possible reason for this may be that students who participate for three years started as freshmen or sophomores, so their capacity for improvement early in their academic career is higher; a single-year participant could be of any class standing, and juniors or seniors may not have as much room for growth over the year due to them having more opportunities to gain forecasting experience outside the WxChallenge.

It is useful to consider all forecasters who participated just one year or multiple years, but another factor in assessing forecasting skill is class standing. By controlling for this variable, which for all intents and purposes means controlling for the amount of formal meteorological education students had through their coursework, with increasing years of experience there was an improvement in mean scores (Table 1), with the exception of seniors participating for the first time, although the sample size for that group was relatively small. By performing a one-way ANOVA test comparing students with different amounts of experience within each class standing, significant differences ($p < 0.001$) were observed among juniors and seniors. Further analysis using a Tukey test revealed

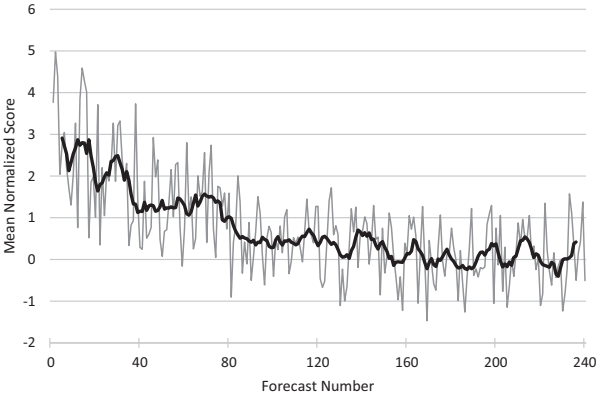

FIG. 6. Mean normalized score by ordinal forecast for participants who submitted 240 human-generated forecasts. The gray curve is the value for each ordinal forecast, and the black curve is the centered 9-day mean value.

TABLE 1. ANOVA test results between number of years of experience in the WxChallenge for each class standing. Significant differences at the 1% level are shown in boldface type.

	Sophomore		Junior		Senior	
Experience (yr)	Count	Mean score	Count	Mean score	Count	Mean score
0	457	5.69	1017	4.47	87	2.36
1	254	4.89	411	4.07	903	3.21
2			106	2.14	229	2.27
3					74	1.29
p value		0.071		**0.000**		**0.000**

significant differences between juniors with two years of experience and juniors with both one year and no experience, each at the 1% level (Table 2). Significant differences only existed between seniors with one year of experience and those with two or three years of experience. For juniors, these results suggest that although improvements are made with more experience, those that competed the two years prior (presumably since their freshman year) had a greater advantage. Interestingly for seniors, the advantage seems to lie in having more than one year of experience. Although the mean scores improved between seniors with two and three years of experience, this difference was not statistically significant. These results support the findings of Roebber and Bosart (1996), who identified experience as the significant factor in forecast skill, as opposed to education.

By holding the number of years of experience that students had in the WxChallenge constant, comparisons between class standings—essentially the amount of meteorological knowledge attained through coursework and forecasting experience from curricular and extracurricular activities other than the WxChallenge—could be assessed. Indeed, among students with no prior participation in the WxChallenge, seniors had the best mean scores, while the mean scores for freshmen were the highest (Table 3). Likewise, for students with one year of experience, scores improved with advanced class standing. A one-way ANOVA test identified significant differences among these two groups of students with similar contest experience levels ($p < 0.001$), but the differences between juniors and seniors with two years of experience was not significant. Using a Tukey test to examine the differences between students with the same amount of experience but different class standings, each of the six combinations of classes with no experience was significant at the 1% level (Table 4).

TABLE 2. The p values from Tukey tests comparing years of experience as based on class standing. The experiment-wise confidence level was 95%. Significant differences at the 1% level are shown in boldface type.

	Juniors	Seniors
0 yr vs 1 yr	0.322	0.261
0 yr vs 2 yr	**0.000**	0.997
0 yr vs 3 yr		0.334
1 yr vs 2 yr	**0.001**	**0.010**
1 yr vs 3 yr		**0.001**
2 yr vs 3 yr		0.272

This shows that while the amount of prior experience a student had in the WxChallenge provides an advantage, knowledge of the atmosphere gained through formal education makes a difference in forecasting performance. A similar result is seen among students with one year of experience, although the differences are only significant between seniors and each of the two classes below them. It is likely that sophomores and juniors have not yet taken many (if any) upper-division meteorology courses prior to the start of the WxChallenge season, so more gains from coursework are realized by the end of a student's junior year. In the context of the findings of Roebber and Bosart (1996), although WxChallenge experience is held constant, it is likely that through additional years of formal education students are gaining experience through exposure to and practice with forecasting products and techniques by way of their courses and extracurricular opportunities.

It is not uncommon for college and university faculty to incorporate required participation in the WxChallenge into their classes, whereby some aspect of a student's participation affects their grade. Illston et al. (2013) surveyed those instructors using the WxChallenge in their classes and found that nearly 90% of them incorporated it as part of the class grade rather than extra credit. Not only have instructors used participation and performance in the WxChallenge as part of the course grade, but they have also tied other activities, such as map briefings and forecast discussions in preparation for WxChallenge forecasts, to the course grade. The mean score for those students who participated through enrollment in a course (6.75) was higher than the students who participated independent of course enrollment (5.08), which was significant at the 1% level using a two-sample t test. This occurred despite the enrolled students actually submitting a greater percentage of possible forecasts than those participating outside a formal class environment (69.6% vs 67.3%). The mean scores of enrolled students were higher than those who participated of their own volition regardless of class standing (Table 5). Two-sample t tests were used to compare the scores between enrolled and nonenrolled students of each class standing, with differences between those who were freshmen, sophomores, and juniors significant at the 1% level, and those who were juniors/seniors (i.e., students who only participated for one year at "category 3") significant at the 5% level. Possible explanations for these results could be found in the motivations underlying the participation by students. While grade motivation has been found to be a strong motivator (Chumbley et al. 2022), which would

TABLE 3. ANOVA test results between class standing as based on number of years of experience in the WxChallenge. Significant differences at the 1% level are shown in boldface type.

	0 years		1 year		2 years	
Class	Count	Mean score	Count	Mean score	Count	Mean score
Freshmen	346	7.35				
Sophomores	457	5.69	254	4.89		
Juniors	1017	4.47	411	4.06	106	2.14
Seniors	87	2.37	903	3.21	229	2.27
p value		**0.000**		**0.000**		0.755

affect those participating through class enrollment, career and intrinsic motivation were also strong motivators. It is not to say that these latter two motivators were not affecting those participating through course enrollment, because for some proportion they likely were, but for those participating without any course grade being involved, one could presume that career and/or intrinsic motivation were likely high. The data do not provide us with any means by which motivation could be assessed, but we can speculate that it played a role in the differences in performance that were observed.

4. Conclusions

This study sought to answer the question, "What trends can we see in 11 years of WxChallenge forecast contest data?" in relation to undergraduate participation in this national weather forecasting contest. Results from an analysis of these data include the following:

- mean scores for students who participated over multiple years showed significant improvement during the first year, but continued improvement through subsequent years was not as clear;
- when controlling for class standing, junior- and senior-level students performed significantly better when they had more than one year of WxChallenge experience;
- after controlling for the number of years of experience, even without prior experience in the WxChallenge, mean scores of students with a higher class standing were significantly better than those with lower class standing;
- the mean scores for those who chose to join the contest on their own were significantly better than those who participated through enrollment in a collegiate course.

TABLE 4. The *p* values from Tukey tests comparing class standing as based on years of experience. The experiment-wise confidence level was 95%. Significant differences at the 1% level are shown in boldface type.

	0 years	1 year
Freshmen vs sophomores	**0.000**	
Freshmen vs juniors	**0.000**	
Freshmen vs seniors	**0.000**	
Sophomores vs juniors	**0.001**	0.059
Sophomores vs seniors	**0.000**	**0.000**
Juniors vs seniors	**0.006**	**0.004**

The results of this study reflect the trends observed by others who studied local forecasting contests, where performance improved through the first 10–30 forecasts (Gedzelman 1978; Hilliker 2008; Suess et al. 2013). However, this study found that while those that participated for just one year improved notably, when the decrease in participation throughout the year was controlled for, the improvements were no longer apparent. By examining the results from a national contest such as the WxChallenge, less is known about those that were participating (e.g., Were they meteorology majors? How many meteorology courses have they taken already? What prior forecasting experience did they bring to the contest?), so interpreting trends such as these can be challenging. And where the trends for single-year participants are less clear, multiyear participants showed improvements through at least the first year, regardless of whether or not participation was controlled for. Likewise, when class standing was controlled for, mean scores generally improved with more years of participation in the contest, reflecting the findings of Roebber and Bosart (1996) regarding the role of experience in forecasting skill.

While Hilliker (2008) was able to ascertain why students did not participate in his contest, we are left to assume that the decline in WxChallenge participation was for similar reasons (i.e., forgetfulness and discouragement). Self-efficacy and motivation are also likely factors in not only explaining the participation trends, especially in single-year forecasters, but also when interpreting the results comparing performance based on participation through enrollment in a class compared to students participating through their own choice. Were career and intrinsic motivation stronger in those students who participated on their own compared to the added grade motivation of those enrolled in a course? Our data do not allow us to know this, but the results clearly indicate that significant differences in performance exist between these two groups, especially for freshmen, sophomores, and juniors.

What these results do not show is whether they are applicable outside the somewhat artificial structure of a forecast contest. The WxChallenge requires participants to forecast specific values for variables (deterministic forecasts), but the jobs that students will one day be pursuing will likely require at least some probabilistic forecasts be made or interpreted (Persson 2013). How well do the skills that students develop in a deterministic forecasting contest translate to probabilistic forecasting? While the answer to this is not known, the results from this study reaffirm the role of experience in cultivating forecasting skill, whether that be through a contest or other activities. Thus, since

TABLE 5. Results from *t* tests between enrolled (registered for a course that required WxChallenge participation) and nonenrolled students by class standing. Significant differences at the 5% and 1% levels are shown with italics and boldface type, respectively.

	Count		Mean score		
Class	Enrolled	Not enrolled	Enrolled	Not enrolled	*p* value
Freshman	120	410	10.52	5.56	**0.000**
Freshman/sophomore	1727	2089	11.32	7.94	**0.000**
Sophomore	312	865	7.23	4.46	**0.000**
Junior	811	1808	4.30	3.74	**0.007**
Junior/senior	4334	3977	6.10	5.38	*0.029*
Senior	859	1407	2.80	2.58	0.144

it is known that the WxChallenge is purely deterministic, inclusion of material and activities related to probabilistic forecasting is recommended for instructors.

There are several opportunities for future work based on this study. All four forecast elements were combined here to produce a singular score, but each could be assessed individually. Precipitation is unique among the forecast elements since it is the only one that frequently does not occur; are WxChallenge forecasters with more experience more skillful in predicting precipitation amounts? Related to the topic of forecast skill, perhaps different choices for the standard of reference—such as climatology, National Weather Service forecasts, or a particular forecast model—might reveal trends or insights beyond those found using the consensus forecast.

Acknowledgments. The authors thank Dr. David Call and four anonymous reviewers for comments and suggestions that improved this paper and Dr. Jeff Basara for his help in obtaining WxChallenge data.

Data availability statement. Data are available from the WxChallenge. Contact the WxChallenge manager for more information (https://www.wxchallenge.com/contact.php).

REFERENCES

AMS Council, 2007: Weather analysis and forecasting: An information statement of the AMS. *Bull. Amer. Meteor. Soc.*, **88**, 1655–1659, https://doi.org/10.1175/1520-0477-88.10.1643.

——, 2015: Weather analysis and forecasting: An information statement of the AMS. Accessed 25 August 2022, https://www.ametsoc.org/ams/index.cfm/about-ams/ams-statements/statements-of-the-ams-in-force/weather-analysis-and-forecasting/.

Bandura, A., 1977: Self-efficacy: Toward a unifying theory of behavioral change. *Psychol. Rev.*, **84**, 191–215, https://doi.org/10.1037/0033-295X.84.2.191.

Barrett, B. S., and J. E. Woods, 2012: Using the amazing atmosphere to foster student learning and interest in meteorology. *Bull. Amer. Meteor. Soc.*, **93**, 315–323, https://doi.org/10.1175/BAMS-D-11-00020.1.

Bloom, B. S., 1956: *Taxonomy of Educational Objectives, Handbook: The Cognitive Domain*. David McKay, 207 pp.

Boi, A. J., and T. C. Spangler, 2012: Addressing complexity in weather: The human role in forecasting. *Earth and Mind II: A Synthesis of Research on Thinking and Learning in the Geosciences*, K. A. Kastens and C. A. Manduca, Eds., Geological Society of America, 113–114, https://doi.org/10.1130/SPE486.

Bonwell, C. C., and J. A. Eison, 1991: Active learning: Creating excitement in the classroom. ASHE-ERIC Higher Education Rep. 1, 121 pp., https://eric.ed.gov/?id=ED336049.

Cervato, C., W. Gallus, P. Boysen, and M. Larsen, 2011: Dynamic weather forecaster: Results of the testing of a collaborative, on-line educational platform for weather forecasting. *Earth Sci. Inf.*, **4**, 181–189, https://doi.org/10.1007/s12145-011-0087-2.

Chumbley, S., M. S. Hainline, T. Wells, and J. C. Haynes, 2022: Students' motivation to learn science through undergraduate-level agricultural coursework. *J. Agric. Educ.*, **63**, 182–199, https://doi.org/10.5032/jae.2022.01182.

Coleman, J. S., and M. Mitchell, 2014: Active learning in the atmospheric science classroom and beyond through high-altitude ballooning. *J. Coll. Sci. Teach.*, **44**, 26–30, https://doi.org/10.2505/4/jcst14_044_02_26.

Decker, S. G., 2012: Development and analysis of a probabilistic forecasting game for meteorology students. *Bull. Amer. Meteor. Soc.*, **93**, 1833–1843, https://doi.org/10.1175/BAMS-D-11-00100.1.

Freeman, S., S. L. Eddy, M. McDonough, M. K. Smith, N. Okoroafor, H. Jordt, and M. P. Wenderoth, 2014: Active learning increases student performance in science, engineering, and mathematics. *Proc. Natl. Acad. Sci. USA*, **111**, 8410–8415, https://doi.org/10.1073/pnas.1319030111.

Gedzelman, S. D., 1978: Forecasting skill of beginners. *Bull. Amer. Meteor. Soc.*, **59**, 1305–1309, https://doi.org/10.1175/1520-0477-59.10.1305.

Grenci, L. M., D. M. Babb, and S. A. Seman, 2008: A grand experiment in e-education: Offering adult students an online apprenticeship in weather forecasting. *Bull. Amer. Meteor. Soc.*, **89**, 969–974, https://doi.org/10.1175/2008BAMS2550.1.

Hamill, T. M., and D. S. Wilks, 1995: A probabilistic forecast contest and the difficulty in assessing short-range forecast uncertainty. *Wea. Forecasting*, **10**, 620–631, https://doi.org/10.1175/1520-0434(1995)010<0620:APFCAT>2.0.CO;2.

Hilliker, J. L., 2008: An assessment of a weather forecasting contest in multi-leveled meteorology classes. *J. Geosci. Educ.*, **56**, 160–165, https://doi.org/10.5408/1089-9995-56.2.160.

Illston, B. G., J. B. Basara, C. Weiss, and M. Voss, 2013: The WxChallenge: Forecasting competition, educational tool, and agent of cultural change. *Bull. Amer. Meteor. Soc.*, **94**, 1501–1506, https://doi.org/10.1175/BAMS-D-11-00112.1.

Knox, J. A., and S. A. Ackerman, 2005: What do introductory meteorology students want to learn? *Bull. Amer. Meteor. Soc.*, **86**, 1431–1435, https://doi.org/10.1175/BAMS-86-10-1431.

Lorditch, E., 2009: Advances in weather analysis and forecasting. *Weatherwise*, **61**, 22–27, https://doi.org/10.3200/WEWI.62.1.22-27.

Morss, R. E., and F. Zhang, 2008: Linking meteorological education to reality. *Bull. Amer. Meteor. Soc.*, **89**, 497–504, https://doi.org/10.1175/BAMS-89-4-497.

Murphy, A. H., 1993: What is a good forecast? An essay on the nature of goodness in weather forecasting. *Wea. Forecasting*, **8**, 281–293, https://doi.org/10.1175/1520-0434(1993)008<0281:WIAGFA>2.0.CO;2.

Persson, A., 2013: The future role of the weather forecaster. *Weather*, **68**, 54–54, https://doi.org/10.1002/wea.2069.

Peyrefitte, A., Jr., and H. Mogil, 1968: A national collegiate forecasting contest. *Weatherwise*, **21**, 162–165, https://doi.org/10.1080/00431672.1968.9932815.

Roebber, P. J., and L. F. Bosart, 1996: The contributions of education and experience to forecast skill. *Wea. Forecasting*, **11**, 21–40, https://doi.org/10.1175/1520-0434(1996)011<0021:TCOEAE>2.0.CO;2.

——, ——, and G. S. Forbes, 1996: Does distance from the forecast site affect skill? *Wea. Forecasting*, **11**, 582–589, https://doi.org/10.1175/1520-0434(1996)011<0582:DDFTFS>2.0.CO;2.

Schultz, D. M., S. Anderson, and R. Seo-Zindy, 2013: Engaging earth- and environmental-science undergraduates through weather discussions and an eLearning weather forecasting contest. *J. Sci. Educ. Technol.*, **22**, 278–286, https://doi.org/10.1007/s10956-012-9392-x.

Skeeter, B. R., 2006: Geography department weather forecasting contests in the 21st century. *J. Geogr.*, **105**, 129–132, https://doi.org/10.1080/00221340608978674.

Stuart, N. A., and Coauthors, 2006: The future of humans in an increasingly automated forecast process. *Bull. Amer. Meteor. Soc.*, **87**, 1497–1502, https://doi.org/10.1175/BAMS-87-11-1497.

Suess, E. J., C. Cervato, W. A. Gallus, and J. M. Hobbs, 2013: Weather forecasting as a learning tool in a large service course: Does practice make perfect? *Wea. Forecasting*, **28**, 762–771, https://doi.org/10.1175/WAF-D-12-00105.1.

Tuan, H. L., C. C. Chin, and S. H. Shieh, 2005: The development of a questionnaire to measure students' motivation towards science learning. *Int. J. Sci. Educ.*, **27**, 639–654, https://doi.org/10.1080/0950069042000323737.

Wilson, K., 2008: Television weathercasters as potentially prominent science communicators. *Public Understanding Sci.*, **17**, 73–87, https://doi.org/10.1177/0963662506065557.

Yarger, D. N., W. A. Gallus Jr., M. Taber, J. P. Boysen, and P. Castleberry, 2000: A forecasting activity for a large introductory meteorology course. *Bull. Amer. Meteor. Soc.*, **81**, 31–39, https://doi.org/10.1175/1520-0477(2000)081<0031:AFAFAL>2.3.CO;2.