Atmospheric Research 100 (2011) 538–546

Contents lists available at ScienceDirect

Atmospheric Research

journal homepage: www.elsevier.com/locate/atmos

Evaluation of European Storm Forecast Experiment (ESTOFEX) forecasts

H.E. Brooks [a,*], P.T. Marsh [b], A.M. Kowaleski [c,1], P. Groenemeijer [d], T.E. Thompson [b], C.S. Schwartz [b,2], C.M. Shafer [b], A. Kolodziej [b], N. Dahl [b], D. Buckey [b]

[a] NOAA/National Severe Storms Laboratory, 120 David L. Boren Blvd., Norman, Oklahoma 73072, USA
[b] School of Meteorology, University of Oklahoma, 120 David L. Boren Blvd., Norman, Oklahoma 73072, USA
[c] Davidson College, Davidson, North Carolina 28035, USA
[d] European Severe Storms Laboratory, Münchner Str. 20, 82234 Wessling, Germany

ARTICLE INFO

Article history:
Received 15 February 2010
Received in revised form 2 September 2010
Accepted 7 September 2010

Keywords:
Severe thunderstorms
Forecasting
Forecast verification

ABSTRACT

Three years of forecasts of lightning and severe thunderstorms from the European Storm Forecast Experiment (ESTOFEX) have been evaluated. The forecasts exhibit higher quality in summer than in winter and there is some evidence that they have improved over the course of the evaluation. Five individual forecasters made the majority of the forecasts and differences in their forecasts are on the order of the overall variability of the forecast quality. As a result, the forecasts appear to come from a single unit, rather than from a group of individuals.

The graphical description of the probability of detection and frequency of hits recently developed by Roebber is a valuable tool for displaying the time series of lightning forecast performance. It also appears that, even though they are not intended for that purpose, using the lightning forecasts as a low-end forecast of severe thunderstorms is potentially useful for decision makers.

Published by Elsevier B.V.

1. Introduction

The European Storm Forecast Experiment (ESTOFEX) was started in 2002 by a group of meteorology students (see http://www.estofex.org/). Its primary goals are to forecast the occurrence of lightning and severe thunderstorm (hail, convective winds, and tornadoes). Although there have been changes over the years in the format of the forecasts, in general, the lightning forecasts have consisted of a line enclosing the area where lightning is expected. The severe thunderstorm forecasts have three levels (1, 2, and 3) of expected coverage and intensity.

Evaluation of forecasts is an important part of the process of improving the forecasts. Besides providing information for the forecasters and users of the forecasts, the ESTOFEX forecasts provide an excellent opportunity to explore the use of relatively new techniques to evaluate and display forecast information.

The question of what makes a good forecast was discussed in an essay by Murphy (1993). He identified three aspects of forecast "goodness." The first is consistency, where the forecasts match the forecaster's "true beliefs." Failures of consistency occur, in large part, because the nature of the forecast system or scores to measure performance may encourage, or even force, a forecaster to forecast something other than his or her expectations. The second is quality, the degree to which forecasts correspond to the observed events. The third is value, measuring the benefits or losses that users obtain by making decisions based on the forecasts. Since we have no access to forecast users and any model that might be constructed of users would be limited in its applicability, we cannot address this aspect in any detail. As a result, our evaluation will be limited almost entirely to the quality of the forecasts.

Several questions are of particular interest to us. First, what changes occur in the forecasts over time? This includes

* Corresponding author. Tel.: +1 405 325 6083; fax: +1 405 325 2316.
E-mail address: harold.brooks@noaa.gov (H.E. Brooks).
[1] Other affiliation: National Weather Center Research Experiences for Undergraduates Program.
[2] Current affiliation: National Center for Atmospheric Research, 3450 Mitchell Lane, Boulder, Colorado 80301.

0169-8095/$ – see front matter. Published by Elsevier B.V.
doi:10.1016/j.atmosres.2010.09.004

the question of secular trends over the course of the forecasting system, but also the seasonal cycle, if any. It has previously been noted that many scores for forecasts of relatively rare events, such as severe thunderstorms, improve as the frequency of the forecast event increases (Doswell et al., 1993). As a result, we might expect forecast quality to improve during times of higher occurrence.

The second question involves the individual forecaster performance. During the period of study, we have five individuals who made almost all of the forecasts. We are interested in whether we can identify those individuals simply by the structure of their forecasts. In other words, would a frequent user of the forecasts be able to tell that the forecast for a particular day was made by a particular individual simply by looking at a graphic that had no name on it? Related to this is the question of whether forecast quality varies significantly between forecasters. In general, we are interested in knowing if the forecasts appear to come from a single forecasting "unit" or from individual forecasters that have distinct characteristics. Many professional forecasting organizations strive to have a degree of uniformity between forecasters, so that users don't have to be aware of which individual prepared the forecast if they intend to make the best use of the product. Although we know, *a priori*, that the individuals are different and that it is likely that detailed study might find differences between the forecasters in narrow aspects, we will be satisfied with looking at gross aspects of forecast differences.

2. Forecast and observational data

The period of analysis runs for 3 years, beginning with forecasts made 30 April 2006, valid for 1 May 2006. The forecast valid time runs for 24 h, beginning at 0600 UTC on the day identified as the forecast day (e.g., 1 May begins at 0600 UTC on 1 May). A sample forecast graphic is shown in Fig. 1. A small number of days had no forecasts issued, resulting in a total of 1038 forecast days. On some occasions, an updated forecast was issued. For consistency, we chose to evaluate only the initial forecast issued for a day. Five forecasters issued 945 (91%) of the forecasts. The smallest number issued by any of that core group was 130. The most issued by the sixth most frequent forecaster was 59. As a result, when we look at individual forecaster performance, we will consider only the core group of five forecasters.

One of the primary requirements for effective forecast evaluation is to match the forecasts and observations. Since the lightning data are gridded, we have put the forecasts and observations onto a grid, so that the events (lightning or severe thunderstorms) are dichotomous and the forecasts are either dichotomous for lightning or ordered (lightning, level 1, 2, or 3) for severe thunderstorms. We will consider each grid point as a forecast-observed pair. The forecasts and observations obviously would have a high degree of spatial correlation, so that even though each point is treated independently, we should not consider them as independent forecasts. The true degrees of freedom within each forecast are much less than the number of forecast points.

Lightning data come from two different sources. Until the end of 2007, the data come from the UK Met Office arrival time difference system. We were provided with information on a $0.5\times0.5^\circ$ latitude–longitude grid every half hour from that system. The information consisted of a scaled value (not total flashes) describing the number of flashes in the time period on the grid. Since the beginning of 2008, lightning data come from the European Cooperation for Lightning Detection (EUCLID—http://www.euclid.org). The format and area of coverage are somewhat different. The spatial grid is $0.25\times0.25^\circ$, but the temporal resolution is 1 h and only part of the ESTOFEX domain is covered.

In order to make the comparison consistent over time, we have put both datasets on a consistent space–time grid ($0.5\times0.5^\circ$, 1 h) using the EUCLID domain (Fig. 2). One or more flashes during the 24-h period for the forecasts are counted as a "yes" event for lightning on that grid.

Severe thunderstorm data come from the European Severe Weather Database (ESWD—http://essl.org/ESWD/) (Dotzek et al., 2009). We have put the reports on the same $0.5\times0.5^\circ$ grid with a 24-h resolution. If one or more reports occur during a forecast valid time in a grid box, the grid box is counted as a "yes" and, if none occur, it's a "no." A significant problem that had to be resolved was the lack of spatial coverage of the ESWD (see parts of the Iberian Peninsula and the Balkans in Fig. 2). We cannot determine, in general, whether the absence of a report is because no weather event occurred or because the reporting system failed to collect the report. (Note that the more mature reporting system in the US makes the latter less common.) We decided to only use those points where severe weather was reported at least once as verification locations for the severe thunderstorm forecasts. In other words, forecasts for locations that never had a severe weather report are not considered in this study. This decision makes sense for locations over water or where reports of real events do not get into the ESWD data base for whatever reason. It is more problematic for locations where severe thunderstorms really didn't occur or for places that received a single report during the 3-year period, but that, in general, events that occurred were not consistently reported. It is impossible to determine the impact of this decision. It does mean that the evaluation is biased towards regions of strong reporting (e.g., Germany). It also means that doing regional verification would be extremely problematic, given the relatively small area that receives reports consistently. As a result, we have chosen not to do any regional comparisons of forecast performance.

3. Methodology

The nature of the two kinds of forecasts leads us to apply different methodologies for their evaluation. At the heart of both techniques is our desire to describe the relationship between forecasts and events, following the general framework of verification of Murphy and Winkler (1987). Since both events are dichotomous (yes or no), our choices are constrained. We can, however, focus on the so-called 2×2 contingency table (see Doswell et al., 1990 for the terminology we will employ), detailing dichotomous forecasts and dichotomous events (Table 1).

For the lightning forecasts, the application of the 2×2 table is straightforward. Both the forecasts and events are inherently dichotomous. We will focus primarily on two quantities, the probability of detection (POD), and the

frequency of hits (FOH). The POD is the fraction of events for which there is a "yes" forecast. The FOH is the fraction of "yes" forecasts for there is a "yes" event. Recently, Roebber (2009) introduced a graphical display that is useful for visualizing those two quantities together and showing the critical success index (CSI) and bias of the forecasts as a function of POD and FOH. As such, it is a natural choice for looking at ESTOFEX's lightning forecasts.

The severe thunderstorm forecasts are inherently ordered forecasts, rather than dichotomous. Ordered forecasts can be summarized graphically via the relative operating characteristics (ROC) diagram, introduced into meteorology by Mason (1982). Brooks (2004) discussed the ROC diagram in the context of the underlying statistical model. It is intended to look at forecast performance when there are forecasts that have a series of ordered levels. Obviously, this is a natural choice for considering the ESTOFEX severe thunderstorm forecasts. It is created by taking each possible forecast level, creating a 2×2 contingency table from it, and then plotting the POD versus the probability of false detection (POFD). The area under a curve (AUC) generated by connecting points at the different forecast levels is a measure of forecast skill and is the Mann–Whitney test statistic. A value of 0.5 represents no skill and a value of ~0.7 is generally considered to be associated with useful forecasts.

One of the questions in the application of the ROC is levels at which a threshold is applied. For a ROC, there are always two default forecasts, always "yes" and always "no." For the always "yes" forecast, the POD = POFD = 1. For the always "no" forecast, the POD = POFD = 0. In between, the order is obvious for the severe level forecasts. Clearly, a level 3 forecast indicates a higher level of threat than a level 2 and a level 2 indicates a

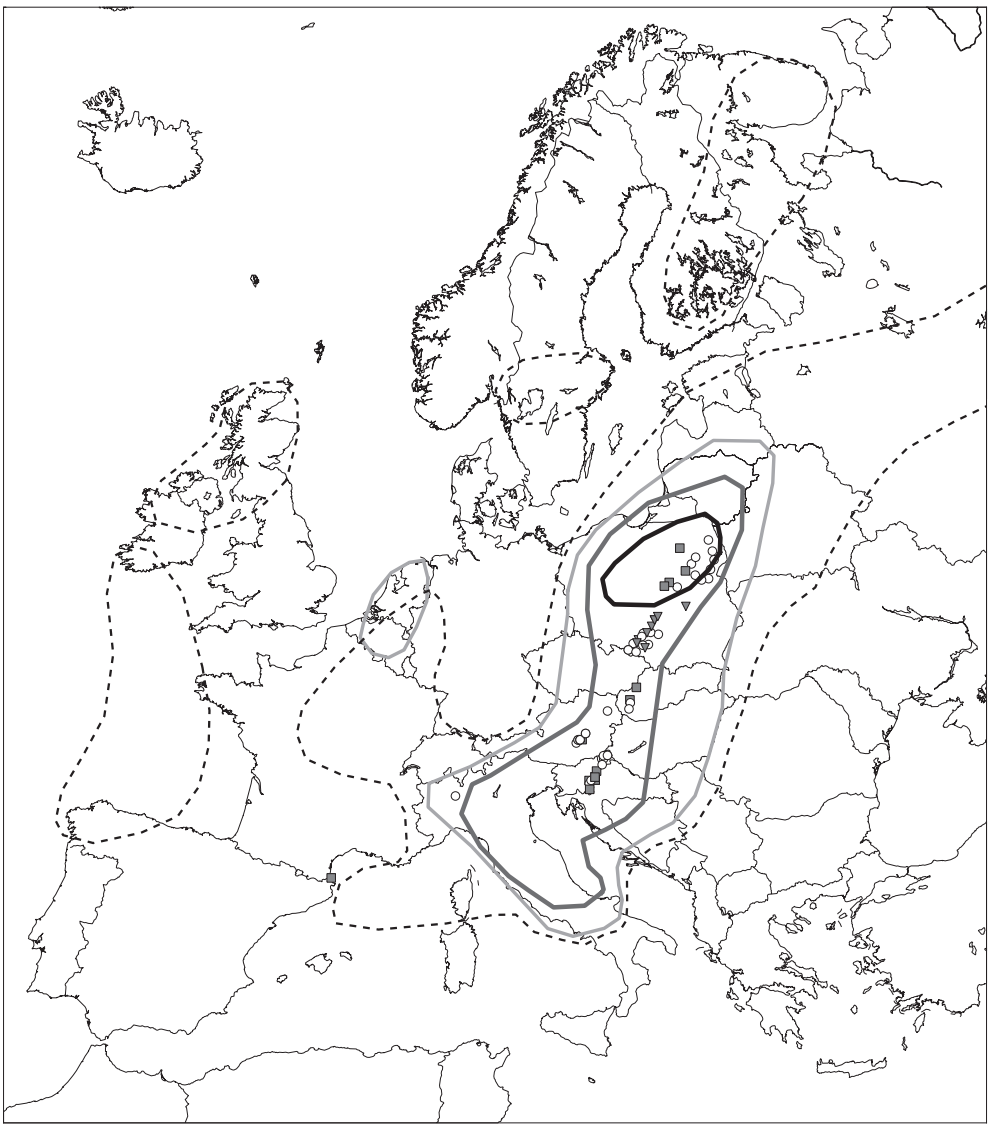

Fig. 1. ESTOFEX forecast issued 14 Aug 2008, valid starting 0600 UTC 15 Aug 2008. Dashed lines indicate regions of expected lightning coverage. Solid lines enclose areas of levels 1, 2, and 3. Observed severe weather reports are shown by symbols (wind-squares, hail-open circles, tornado-inverted triangles.

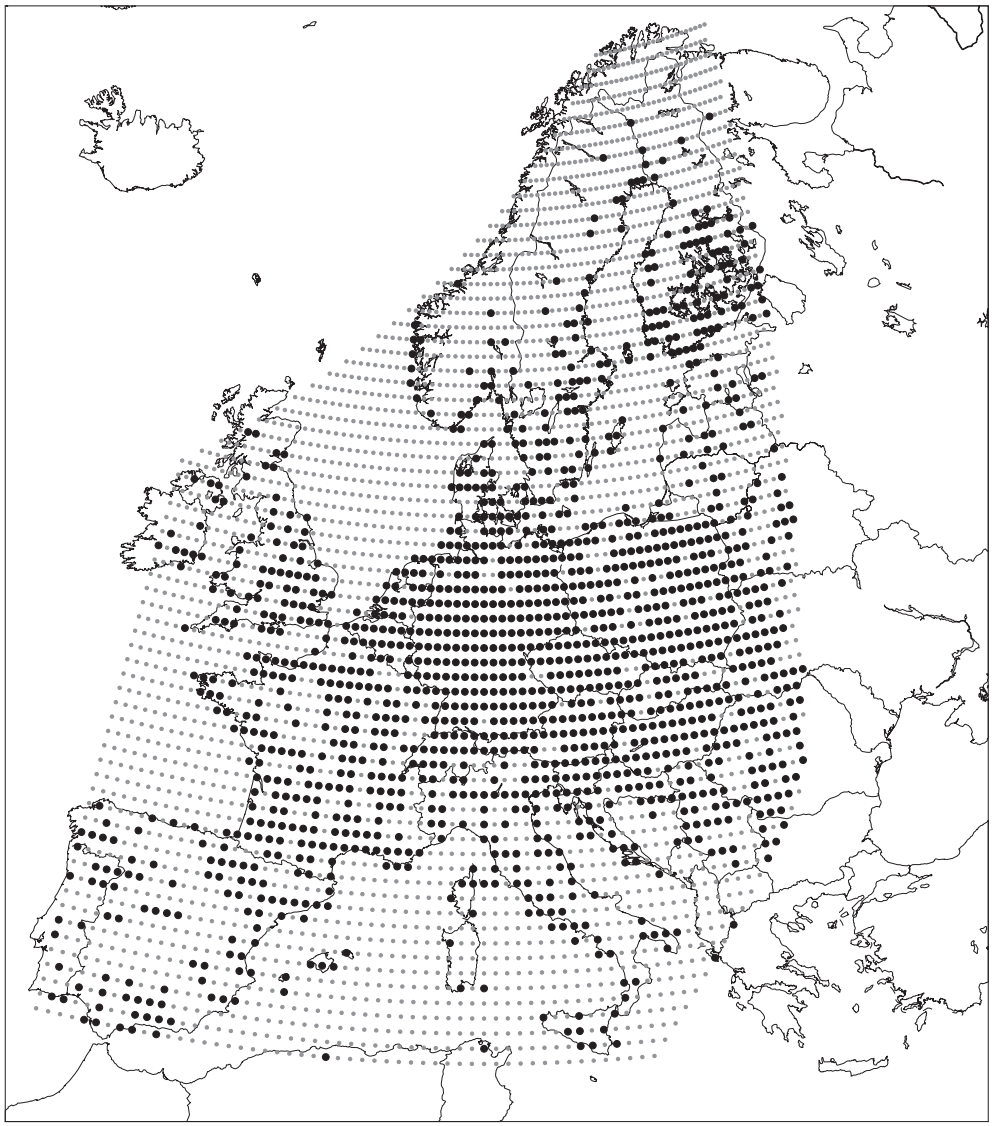

Fig. 2. Verification locations for forecasts. Small gray dots represent lightning verification locations. Large black dots are those locations where severe thunderstorms were reported at least once during the verification period.

higher level than level 1. A question arises about what to do with lightning as a forecast for severe thunderstorms. Typically, the lightning forecast area encloses the severe forecast area, but that is not required. There are some cases in which severe forecasts are found outside of the lightning area, but those are relatively rare. We have chosen to use the lightning forecast as a severe thunderstorm forecast level of level 0. For individual forecasts, this can be problematic, since there may not be a lightning forecast, but for analysis of groups of forecasts, the ambiguity in interpretation should be small.

Table 1
2×2 contingency table for forecasts and observations.

	Observed yes	Observed no	Sum
Forecast yes	a	b	a + b
Forecast no	c	d	c + d
Sum	a + c	b + d	n

Quantities of interest: Probability of detection (POD) = a/(a + c). Frequency of hits (FOH) = a/(a + b). Probability of false detection (POFD) = b/(b + d). Critical success index (CSI) = a/(a + b + c). Bias = (a + b)/(a + c).

4. Results

4.1. Lightning forecasts

We begin by considering the relationship between the forecast and observed areas, without regard for the relationship between individual points on those forecasts. For ease of interpretation, we have averaged the values over 91 consecutive forecasts. This produces something that looks like seasonal averages, without imposing an arbitrary calendar on those seasons. There is a strong annual cycle to both the observed and

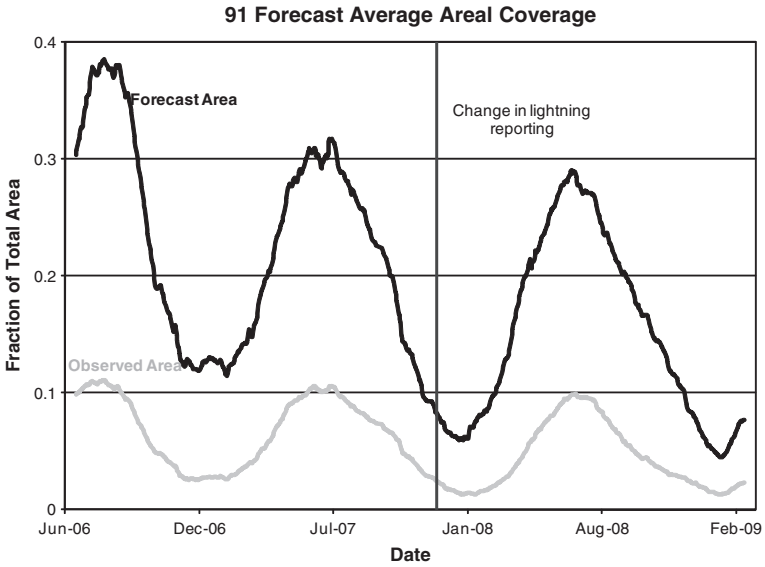

Fig. 3. Fraction of lightning domain with observed (gray line) and forecast (black line) coverage of lightning in 24-hour period. Lines are 91-day running means. Vertical line indicates change from UK Met Office to EUCLID lightning data.

forecast lightning areas (Fig. 3), with the peak in summer when approximately 10% of the area has lightning. There is a slight decrease from year to year in the observed coverage. Clearly, the lightning area is overforecast throughout the three years. The forecast area decreases more over the three years than the observed, which may indicate that the forecasters were becoming calibrated, but the forecast area is three times the observed area in the warm season. Overforecasting of rare events is not necessarily a bad thing, if the cost of a missed event is greater than the cost of a false alarm. It arises in a number of contexts and can be a consequence of a desire to have a relatively high POD (Brooks, 2004).

One simple way of looking at individual forecasters' output is to look at the size of each forecast. This addresses the question of whether some forecasters make larger forecasts than others. A useful way of showing this is to use a probability of exceedance diagram. Here, the probability that a forecast will be equal to or exceed some threshold size is shown compared with the threshold. Obviously, all forecasts will be at least as large as the smallest forecast ($p=1$) and the probability of exceeding larger thresholds will monotonically decrease as the threshold increases. The curves showing the probability of exceedance for the five core forecasters shows relatively small differences between them

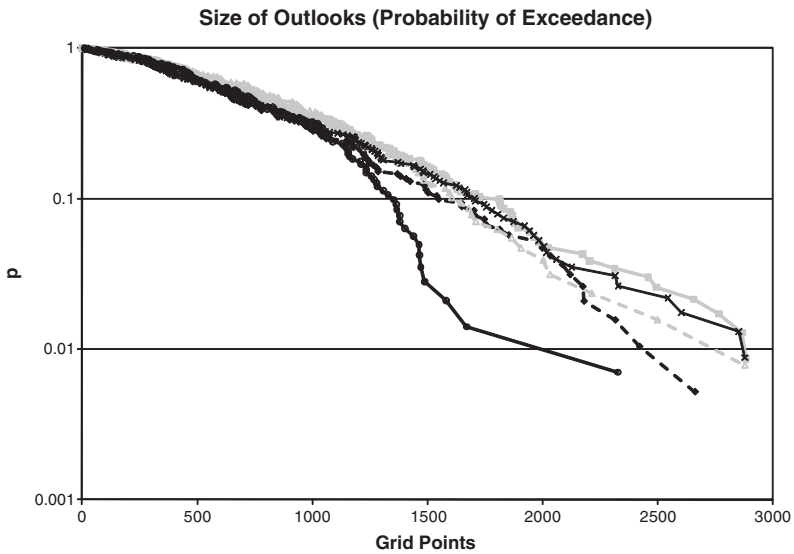

Fig. 4. Probability of exceedance of forecast size for each of five core forecasters. Lines show probability (vertical axis-log scale) that a forecast will be at least as large as the number of grid points on the horizontal axis. Each line represents a different (anonymous) forecaster.

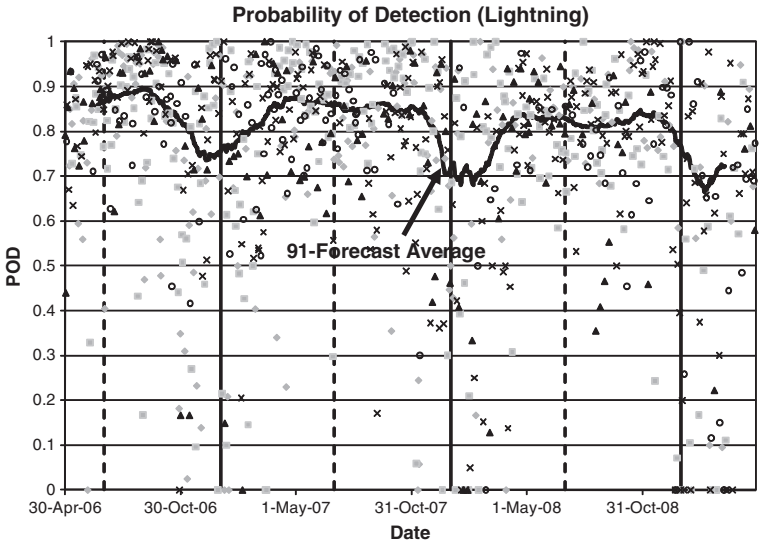

Fig. 5. Probability of detection for lightning forecasts. Symbols represent daily values (each symbol for a different forecaster) and solid line is 91-forecast performance centered on date on horizontal axis. Solid vertical lines at 1 January, dashed vertical lines at 1 July.

except for the largest forecast areas (Fig. 4) (note that the areas shown are for the entire ESTOFEX forecast domain, not just the domain for lightning verification.) The differences in the largest forecast areas, corresponding to the largest 10% of individual forecasts, could be a result of differences in when forecasters worked. It is highly unlikely for a forecaster who makes most of their forecasts in winter to have very large areas. As a result, it is difficult to distinguish forecasters by the size of the forecasts.

As a starting point for looking at the quality of the lightning forecasts, we consider the POD of the forecasts (Fig. 5). This illustrates the impact of the 91-forecast averaging. It also shows that the low POD time periods result from a larger fraction of low POD forecasts in the cold season. A more complete picture of performance can be drawn from the Roebber diagram (Fig. 6). The forecasts tend to be in the high POD-low FOH part of the phase space, where bias is much greater than 1, as seen in the areal coverage figure in Fig. 3. The forecasts are far from perfect (POD = FOH = 1) in the upper right hand corner of the diagram, but much better than would be expected from a purely random combination of the forecasts and observations (POD = 0.18, FOH = 0.07). In fact, the random forecast is so easy to outperform that it is effectively meaningless.

As before, we can smooth over 91-forecast periods to make the progression through time easier to see. Calculating the POD and FOH over 91-forecast periods and then running those time series through a local regression with a light smoother[3] produces a visually attractive plot (Fig. 7). Since the region of interest is only a small part of the domain, we can focus on the upper left portion of the diagram (Fig. 8). The forecasts are clearly of higher quality in the summer than in the winter, with the CSI approximately 0.1 higher in the warm season. In general, when events are more frequent, the CSI is higher. The 3-year period is relatively short to identify trends or even a consistent annual cycle, outside of the summer/winter difference. Qualitatively, the CSI doesn't change much from year to year, but the forecasts have moved towards lower values of bias (towards the lower right.) Given that the changes are relatively small for the other measures, this is a desirable change.

The overall performances of the individual forecasters can also be compared with the trend. They forecasters are all in the upper right of the 3-year record. Although this may look paradoxical, the values of the POD and FOH in the overall performance are dominated by "big days." In the limit, if there were a large number of days with only one point with a yes forecast or event, and then one day with hundreds of yes

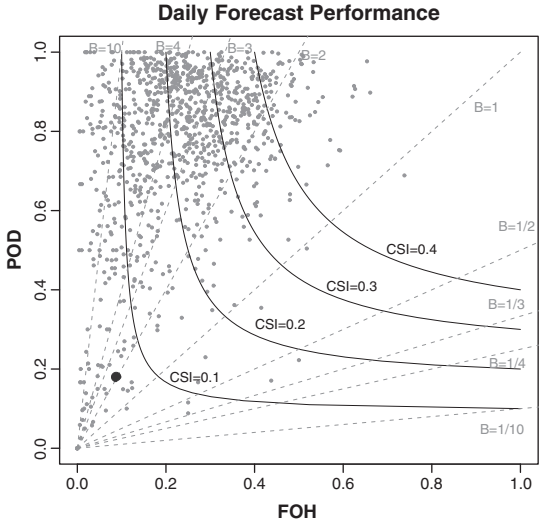

Fig. 6. Roebber (2009) diagram of forecast performance in terms of probability of detection (POD) and frequency of hits (FOH). Each small dot represents one day's forecast and the large black dot (FOH = 0.07, POD = 0.018) is the performance of a random forecast. Dashed lines indicate bias (B) and curved lines are critical success index (CSI). Perfect forecasts in upper right (POD = FOH = 1).

[3] The smoother used is the LOWESS routine in the statistical package R, with a smoothing parameter of $f = 0.05$. R is available at http://cran.r-project.org.

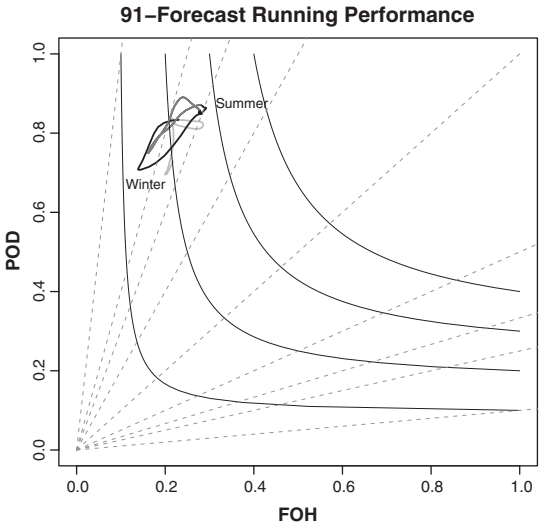

Fig. 7. Same as Fig. 6, except smoothed 91-forecast running mean. Black line is first year of experiment, medium gray second year, and light gray third year. "Summer" and "Winter" indicate relative positions of performance on annual cycle.

points, the overall performance would be much closer to the POD and FOH of the large forecast/event day than it is to the 0 or 1 from the "one-point" days. Resampling the 1038 forecasts to see the inherent variability of the performance indicates that the 95% confidence interval extends to just to the left of the highest FOH individual forecaster and includes the other four points. It is important to note, however, that this forecaster had the larger fraction of warm-season forecasts of any of the group. A simple resampling approach is inadequate to estimate the variability of performance. In the absence of an objective measure of forecast difficulty, it is not clear how to carry out meaningful statistical hypothesis tests. It is clear that the distribution of forecasts was far from random. It is possible that the relationship that we have assumed (forecast performance is better in warm season, so that forecasters who have more of their forecasts in the warm season will appear to be better) could be backwards. It could be that the reason that the warm season looks better is because the better forecasters work more often then. This seems unlikely, however, given preliminary results from a 30-year analysis of severe thunderstorm forecasts from the United States, that also show better performance in the warm season with a larger sample size and a more uniform distribution through the year of individual forecasts (R. Lam, 2010, personal communication).

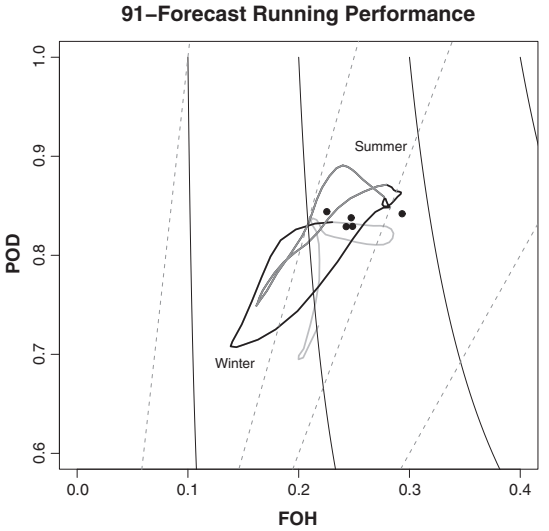

Fig. 8. Zoomed-in portion of Fig. 7, with black dots indicate overall performance by individual forecasters.

4.2. Severe thunderstorm forecasts

The structure of the severe thunderstorm forecasts, with the different forecast levels and dichotomous events, lends itself to analysis via the ROC diagram. As an example, we look at the overall forecast performance for the entire dataset (Fig. 9). The lower left-hand corner of the curve (POD = POFD = 0) is associated with the default forecast that the event (severe thunderstorms) will never occur. There is another point very near the (0,0) location associated with level 3 forecasts (POD = 0.004, POFD = 0.00003). These forecasts are used so rarely that very few severe weather events occur within them, but they also have very few false alarms. Moving up the curve towards the upper right corresponds to going to lower forecast levels. By including the lightning forecast as a forecast for severe thunderstorms, an additional point can be added to the curve created just from the severe thunderstorm forecasts. The area under the curve (AUC) is a measure of the ability of the forecast system to discriminate between when severe thunderstorms occur and when they don't. The area with just the severe thunderstorm forecasts included is 0.72, but with the lightning forecasts included, it increases to 0.86. Given that discriminators associated with AUC ~ 0.7 are considered useful, the severe thunderstorm forecasts can be considered useful, but the

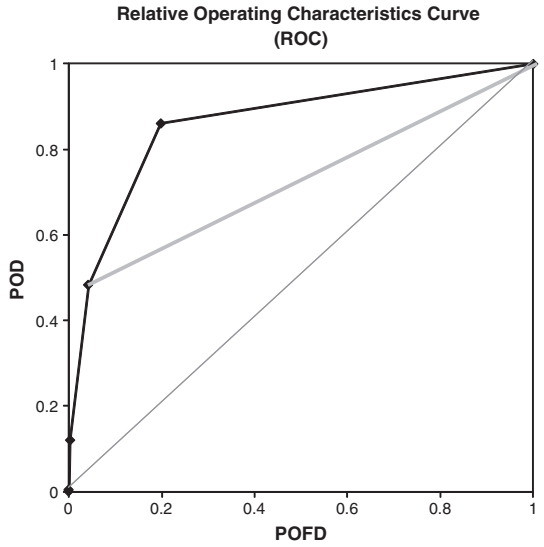

Fig. 9. Relative operating characteristics curves for overall severe thunderstorm forecast performance. Solid black line is associated with severe thunderstorm forecasts with lightning forecast included as lowest level of forecast (AUC = 0.86). Thick gray line is associated with not including lightning forecast as lowest forecast (AUC = 0.72). Thin gray line along diagonal (POD = POFD) represents no-skill forecast.

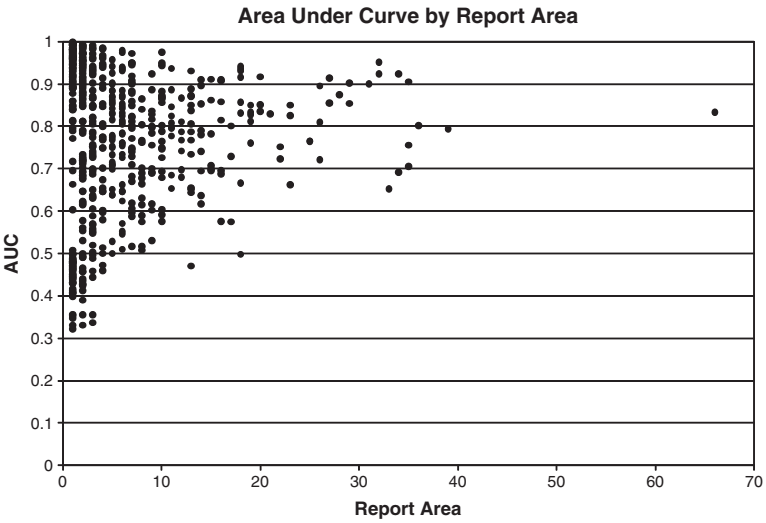

Fig. 10. Area under ROC curve for individual forecasts as function of number of grid points with severe thunderstorm reports for the day.

inclusion of the lightning forecasts dramatically increase the utility.

One way to consider the different points on the curve is to look at them from the perspective of a decision maker. The complete curve describes the forecast system (in this case, the ESTOFEX forecasters), but the individual points represent thresholds at which decision makers might decide to take action or not. Users who are sensitive to missed events might decide to take action at a lower threshold of confidence (higher up the ROC diagram), while those who are relatively more sensitive to false alarms would require a higher threshold (lower on the ROC diagram). Thus, the AUC tells us something about overall performance, but the individual points are useful for particular users.

The AUC can be calculated for any particular forecast day. The number of grid boxes on a day with severe thunderstorm reports has a strong effect on the AUC for individual days (calculated including the lightning forecast) (Fig. 10). For small areas, the AUC can take on almost any value, including on some days, values less than 0.5, indicating a performance worse than random. If the area is greater than about 10 grid points, the range of observed AUCs narrows considerably. For the roughly 30 forecast days with reports in more than 20 grid boxes, the range narrows even more, ranging from ~0.65 to ~0.95. The cluster of values above 0.8 leads to the overall value of 0.86 seen in Fig. 8.

The annual cycle of performance can be seen by plotting the AUC for every forecast (Fig. 11). In winter, with infrequent severe thunderstorm occurrence, values are either near 1 or 0.5, indicating that the forecaster either got a rare event in a small forecast area or missed it. In the summer, with the larger forecast areas and event coverage, a broader range of

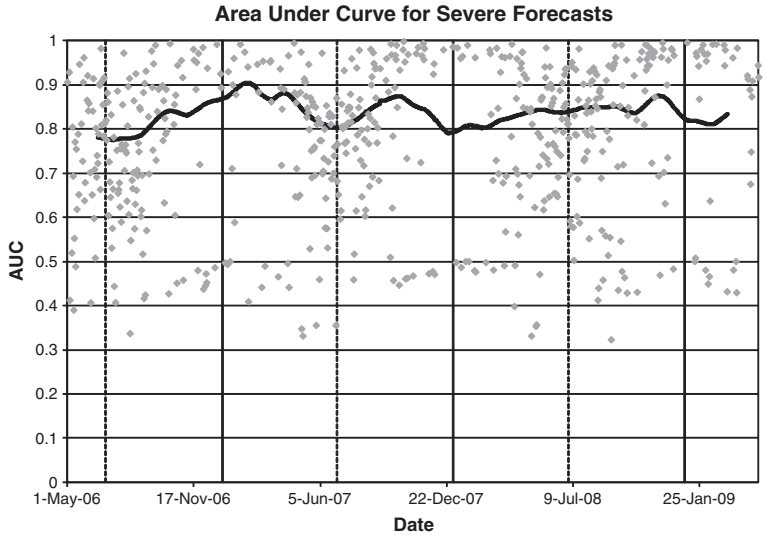

Fig. 11. Area under ROC curve for severe thunderstorm forecasts, including lightning forecast as a forecast for severe thunderstorms. Dots represent daily forecasts, solid line is 91-forecast forecast performance, centered on date on horizontal axis. Solid vertical lines at 1 January, dashed vertical lines at 1 July.

values is seen. Again, it is more informative to calculate values from a set of 91 consecutive forecasts to seeing the long-term changes in forecast performance. In contrast to the lightning forecasts, in general there is a long-term increase in warm-season forecast performance, but the seasonal signal is not very consistent. The average forecast is useful, in terms of the AUC, almost all of the time. Despite the 91-forecast averaging, small sample size issues still exist. The abrupt change in January 2008 results from a single high-quality large area forecast of many events becoming a part of the averaging window following a quiet period of a couple of months.

5. Discussion and conclusions

The ESTOFEX forecasts are of a reasonably good quality and there is evidence that differences in forecaster performance are on the order of or smaller than variability in forecast difficulty, so that the simple analysis here cannot distinguish between the five forecasters. Forecasts are better in the warm season and there's weak evidence that lightning forecasts have improved from year to year and somewhat stronger evidence from the AUC calculations that the severe thunderstorm forecasts have improved.

The primary barrier to doing more complete analysis of the severe thunderstorm forecasts is the lack of reporting of events throughout the region. As a result, the forecast performance is likely biased towards central Europe. It's possible that performance in other regions is different from what is shown here. This also means that follow-up analyses focusing on problem areas in the forecast system could be misplaced. If we are to have more confidence in the ESTOFEX forecasts and to trace changes in their quality through time, the reporting database has to improve and, ideally, reports arrive in near real-time to give forecasters rapid feedback.

Acknowledgments

AMK participated in the 2009 National Weather Center Research Experiences for Undergraduates, supported by the National Science Foundation under Grant No. ATM-0648566. We thank the EUCLID network for kindly providing lightning detection data. We thank Richard Lam of the University of Oklahoma School of Meteorology for sharing the results of his analysis of the Storm Prediction Center forecasts.

References

Brooks, H.E., 2004. Tornado warning performance in the past and future: a perspective from signal detection theory. Bull. Amer. Meteorol. Soc. 85, 837–843.

Doswell III, C.A., Davies-Jones, R., Keller, D.L., 1990. On summary measures of skill in rare event forecasting based on contingency tables. Weather Forecasting 5, 576–585.

Doswell III, C.A., Johns, R.H., Weiss, S.J., 1993. Tornado forecasting: a review. In: Church, C., Burgess, D., Doswell, C., Davies-Jones, R. (Eds.), The tornado: its structure, dynamics, hazards, and prediction. Geophys. Monogr., 79. Amer. Geophys. Union, pp. 557–571.

Dotzek, N., Groenemeijer, P., Feuerstein, B., Holzer, A., 2009. Overview of ESSL's severe convective storms research using the European Severe Weather Database ESWD. Atmos. Res. 93, 575–586.

Mason, I., 1982. A model for assessment of weather forecasts. Aust. Meteorol. Mag. 30, 291–303.

Murphy, A.H., 1993. What is a good forecast? An essay on the nature of goodness in weather forecasting. Weather Forecasting 8, 281–293.

Murphy, A.H., Winkler, R.L., 1987. A general framework for forecast verification. Mon. Weather Rev. 115, 1330–1338.

Roebber, P.J., 2009. Visualizing multiple measures of forecast quality. Weather Forecasting 24, 601–608.