Verification of ESTOFEX Lightning and Severe Weather Forecasts

Alex Kowaleski
Davidson College

Dr. Harold E. Brooks
NOAA/National Severe Storms Laboratory

Dr. Charles A. Doswell III
Cooperative Institute for Mesoscale Meteorological Studies

Abstract

The European Storm Forecast Experiment's (ESTOFEX) daily 2006-2009 ordered lightning and severe weather forecasts were analyzed by using a two by two contingency table. Probability of detection (POD), frequency of hits (FOH), probability of false detection (POFD), critical success index (CSI), and bias were calculated. These scores were compared among seasons and years to determine how forecasting skill varied by season, and how it changed from 2006 to 2009. They were also compared among forecasters to determine if some forecasters were more skilled than others. It was determined that severe weather forecasts improved in both POD and FOH scores between 2006 and 2009, Forecasts of lightning, however, did not consistently improve during the forecasting period. It was also determined that ESTOFEX issued superior lightning and severe weather forecasts during summer, and their forecasts were less successful during fall and winter. The differences in forecasting success among forecasters, however, were not sufficiently large to determine if some forecasters were more skilled than others. An analysis of the Relative Operating Characteristics curves (ROCs) of ESTOFEX severe weather forecasts indicated that they were useful for decision-making.

1. Introduction

Although severe convective thunderstorms cause 7 to 11 billion dollars of damage a year in Europe (Dotzek, et al., 2009), there is no continent-wide authority that forecasts these storms, as each nation is largely responsible for forecasts within its borders. Starting in 2002, however, a group of young European meteorologists formed the European Storm Forecast Experiment (ESTOFEX), an effort to predict lightning and severe weather throughout Europe by issuing daily forecasts similar to those that the United States' Storm Prediction Center issues.

Each day at 12Z an ESTOFEX meteorologist issues an ordered lightning and severe weather forecast for the following day in Europe. Lightning forecasts are categorically divided into "lightning" and "no lightning" forecast areas. ESTOFEX divides its severe forecasts into severe level one through severe level three thresholds. Increasing thresholds represent forecasts of increasing probability and/or severity of severe storms. ESTOFEX defines a severe convective storm as a storm with lightning that produces one or more of the following: wind gusts at least 25 meters/second, hail at least 2.0 centimeters in diameter, or a tornado.

Much of the verification of severe weather in the United States has focused on the probability of detection (POD), frequency of hits (FOH), and critical success index (CSI) (Brooks, 2004). Because a missed severe weather forecast causes greater harm than a false alarm,

Corresponding author address: Alex Kowaleski
1626 Eagle Nest Circle Winter Springs, FL 32708

priority is given to increasing the POD, though this can increase the number of false alarms, decreasing the FOH. A successful severe weather forecast therefore maximizes the POD while keeping the FOH as high as possible. The CSI, a function of the POD and FOH, provides an overall measure of forecasting success.

The POD, FOH, and CSI, as well as the probability of false detection (POFD) can be calculated by the creation of a two by two contingency table (Donaldson et al., 1975) (Doswell et al, 1990). For ESTOFEX's lightning forecasts, only one two by two contingency table is required. However, for the severe weather forecasts, four tables are needed, due to the four ordered levels of threat (lightning-severe 3). In each table, the determined. This allows calculation of the POD, FOH, CSI, POFD, and bias.

2. Methodology

To verify ESTOFEX's lightning forecasts, Europe was divided into ½ by ½ degree grids, and a point was placed at the center of each grid box (Fig. 1). This created 4047 individual point forecasts to verify each day. ESTOFEX issues forecasts beyond this grid, but the quality of lightning reporting from outside the grid is poor. For severe weather verification, points on the grid that reported no severe weather during the forecast period were excluded. This left a total of 1392 points. Because Europe's severe weather reporting network is inferior to that of the United States, it is impossible to determine whether these areas did not receive severe weather, or whether they experienced severe weather but it went unreported. Although this filtering was necessary, it excluded large areas of Europe, such as parts of the Iberian Peninsula, southern Italy, and the western Balkans, while including nearly all points in central Europe in the severe forecast verification area.

Each of ESTOFEX's forecast days between April 30, 2006 and April 30, 2009 was verified, a total of 1038 forecast days out of 1097 days during the period. For each day, it was determined which of the 4047 lightning reporting points fell within areas where lightning was forecasted, and which of the 1392 severe weather reporting points fell within the lightning and severe one through severe three forecast areas.

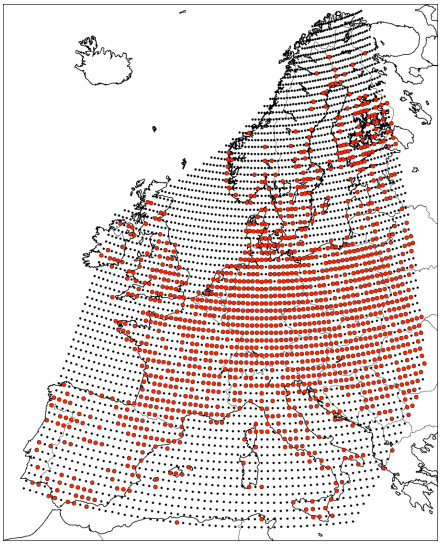

Fig 1. Map of Europe used for ESTOFEX verification. All 4047 points were each used for lightning verification. Only the larger 1392 points were used for severe weather verification, as these were the locations where severe weather had been reported during the three year verification period. The relative absence of severe weather reporting in parts of southern Europe is apparent.

Severe weather reports were obtained from the European Severe Weather Database, a division of the European Severe Storms Laboratory. Lightning reports were obtained from the United Kingdom Meteorological Office, which records lightning strikes throughout Europe.

After determining the total number of points each day where ESTOFEX issued lightning and severe level one through severe level three forecasts and the total numbers of points inside and outside the forecast areas where lightning and severe weather occurred, the two by two contingency table was used to calculate the POD, FOH, CSI, POFD, and bias. Multiple forecasts were combined to calculate totals by year, season and forecaster. Scores were also averaged over a 91 day forecast period to determine how forecasting success changed with time. A 91 day forecast average was chosen because it represented the length of a season, while showing continuous change.

Roebber diagrams, which Paul Roebber introduced in *Weather and Forecasting* (Roebber, 2009) were created for lighting and severe weather forecasts, allowing geometric comparison of forecasts by year, forecaster, and season.

Relative operating characteristic curves (ROCs) were also used to determine the success of severe thunderstorm forecasts. The ROCs, which measure forecast usefulness for decision making, were created by plotting the each threshold's forecast on a POFD-POD axis. These points were then connected, as well as connected to trivial forecasts at (0, 0) and (1, 1). (0, 0) and (1, 1) represent theoretical thresholds in which severe weather is never and always predicted, respectively.

3. Lightning Verification Results

The size of forecasted and observed lightning areas followed an annual cycle, peaking in summer and reaching their lowest value in winter (Fig 2). During all parts of the forecast period the bias was greater than one, as a larger area was predicted to have lighting than the area where lightning occurred. However, during the three years of forecasting, the size of the forecast area decreased, especially between year one and year two, which resulted in a higher FOH score because of the reduced number of false alarms.

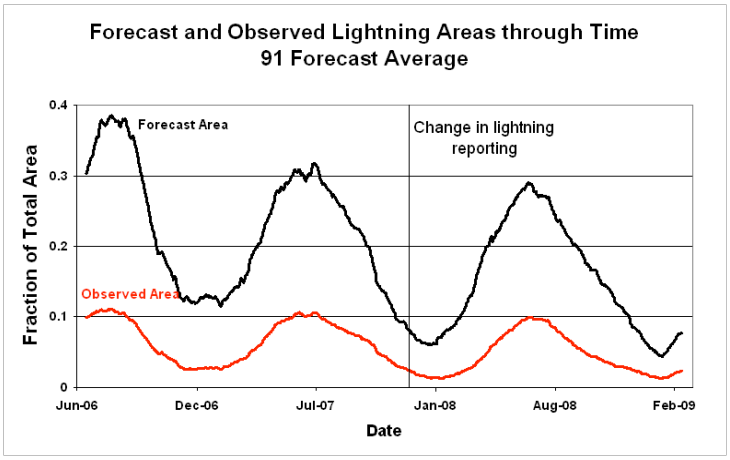

Fig. 2. 91 forecast-day averages of forecasted and observed lightning areas as functions of date. Both forecast and observed areas were highest during the summer and lowest during the winter. Forecast area decreased each year, especially between year one and year two, leading to less false alarms and a higher frequency of hits (FOH score).

Forecast success, as measured by POD, FOH, and CSI also followed an annual cycle, reaching their highest values during the summer and their lowest during the winter. The POD, FOH and CSI improved markedly from year one to year two, before declining between year two and year three (Fig 3).

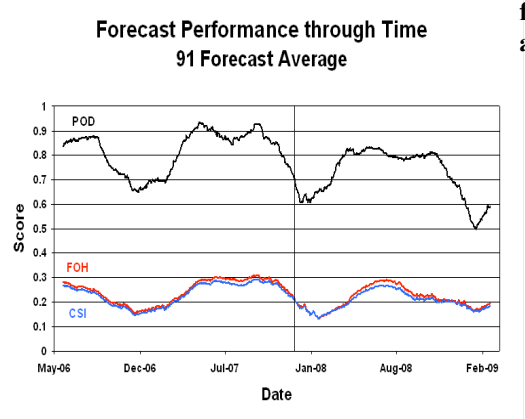

Fig. 3. 91 forecast-day averages of probability of detection (POD), frequency of hits (FOH), and critical success index (CSI) as functions of date. Scores were best during the summer months and worst during the winter months. Scores improved between year one and two, but decreased between year two and year three.

The Roebber diagrams (Figs. 4 and 5) which compares forecasts by forecaster, season and year, shows that in the aggregate, forecasts improved in CSI from 0.23 to 0.25 between year one and year two. Year three's aggregate CSI, however, fell slightly to a value of 0.24. Overall, the distribution in CSI among years was much smaller than the distribution among seasons and among forecasters.

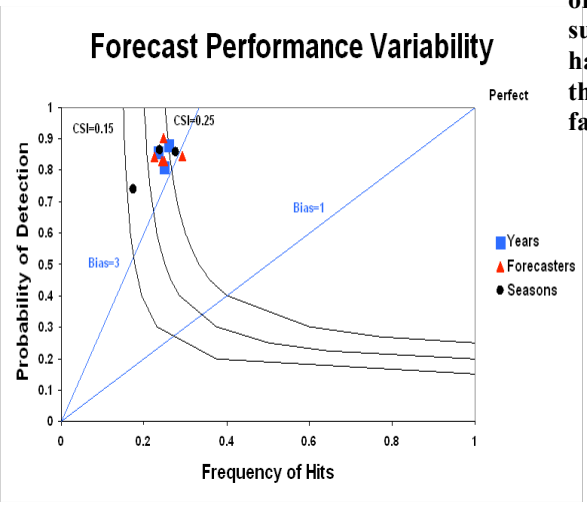

Fig 4. Adaptation of a diagram developed by Paul Roebber. It shows a geometric comparison of ESTOFEX's lightning forecasts by year, forecaster, and season. A perfect forecast would appear in the upper right-hand corner.

Forecasting success among seasons varied greatly, with summer lightning forecasts averaging a CSI of 0.26, while winter forecasts averaged 0.16. Spring and fall forecasts obtained CSI scores of 0.23. Lightning forecasting was therefore most successful during summer and least successful during winter. Success among forecasters varied less than among seasons; the highest average CSI of a forecaster was 0.28 and the lowest average was 0.22.

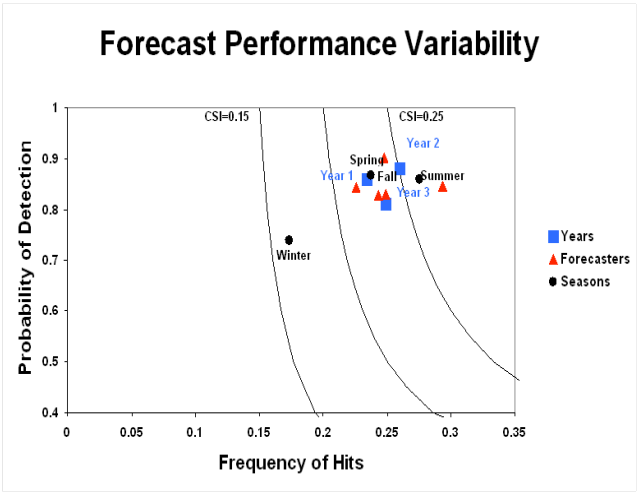

Fig 5. Top left-hand corner of Roebber diagram of lightning forecasts. Year two had the most successful forecasts, with years one and three having lower CSI scores. Summer forecasts were the most successful, and winter forecasts were by far the least successful

The Roebber diagram of 91 forecast-day averages (Fig. 6) shows the progression of forecasts throughout each year and between years. CSI scores in each of the three winters were comparable; while the winter of 2008-2009 had the lowest POD, its CSI score was balanced by a higher FOH. Year two's greater forecast success was primarily due to the superior forecasting performance in the summer of 2007, which reached higher CSI scores and remained at the high scores for more days

than year one or year two's summer (2006 and 2008, respectively).

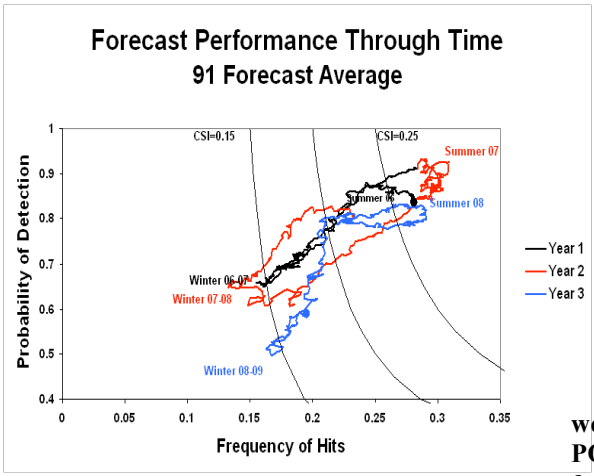

Fig. 6. Roebber diagram of changes in POD and FOH over time. During each winter, the CSI scores were nearly the same, although the POD and FOH varied. Year two had a higher CSI score than year one or three because the summer of 2007 had a large number of forecasts with a high POD and FOH.

4. Severe Verification Results

Like lightning forecasts, severe weather forecasts achieved their highest CSI scores during summer. However, unlike lightning, severe weather forecasts earned their lowest scores during fall, rather than winter. Also unlike lightning forecasts, severe weather forecasts showed consistent improvement between year one and year three, as the CSI increased from 0.018 in year one to 0.023 in year two and 0.026 in year three.

Variation in CSI among forecasters was comparable to variation of CSI among seasons, with seasonal averages ranging from 0.011 to 0.026, and forecaster averages ranging from 0.018 to 0.032 (Fig. 7).

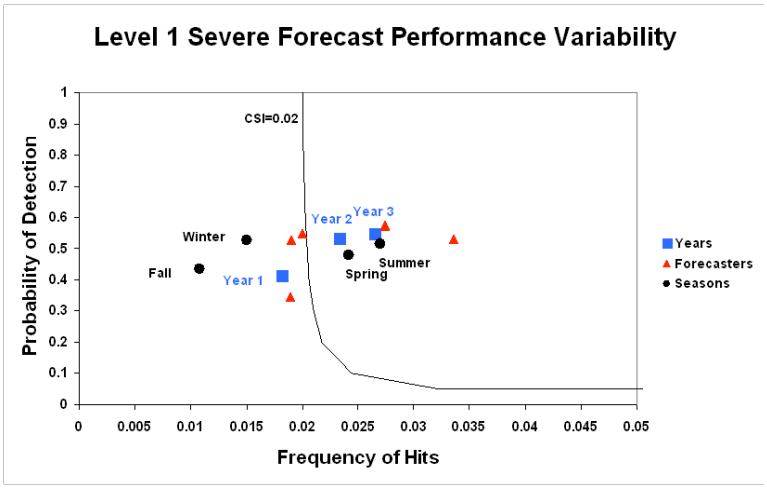

Fig. 7. Roebber diagram for level 1 severe weather forecasts. Forecasts improved in both POD and FOH between each year. Like lightning forecasts, the severe weather forecasts were best in the summer, but unlike lightning, the worst severe weather forecasts occurred during the fall rather than the winter.

The overall Relative Operating Characteristic (ROC) curve diagram for the average of all severe weather forecasts yielded an area under the ROC curve of 0.86 (Fig 8). Numerous studies in meteorology, medicine, psychology, and other scientific disciplines have shown that for a forecast to be useful, the area under the ROC curve must be at least 0.7.

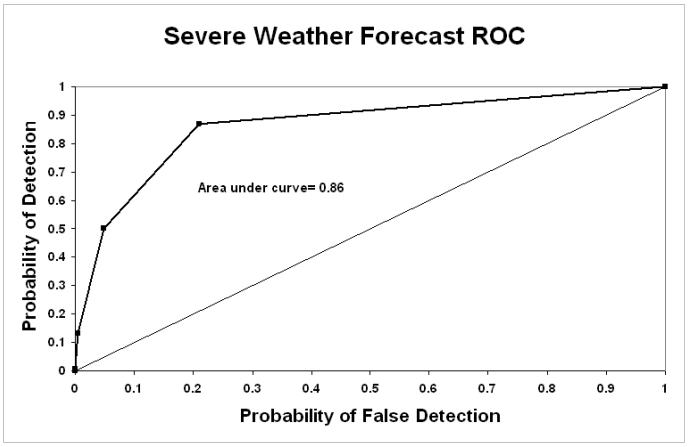

Fig. 8. Relative Operating Characteristics (ROC) curve for the average of all ESTOFEX severe weather forecasts. A forecast is useful if the area under the ROC curve is greater than or equal to 0.7.

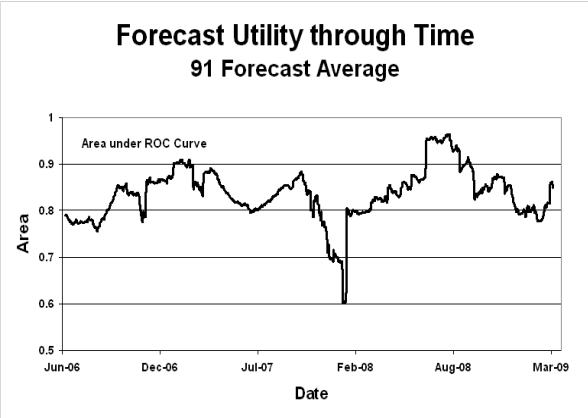

Fig. 10. Area under the Relative Operating Characteristics curve (ROC) curve as a function of date. For almost the entire forecast period, except for a short time in early 2008, the area under the ROC curve was greater than 0.7, indicating that ESTOFEX's severe weather forecasts were useful for decision making.

Therefore, ESTOFEX's severe weather forecasts were useful in the aggregate. The 91 forecast-day running average of ROC scores showed that the usefulness of forecasts varied from a peak area under the curve of 0.96 to a low area of 0.60 (Fig 9). Therefore, except for a short period of time in early 2008 when the 91 forecast average dipped below 0.7, ESTOFEX's severe weather forecasts are useful for decision making.

5. Summary and Conclusion

ESTOFEX's lightning and severe weather forecasts were most successful during summer. Lightning forecasts were least successful during winter, and severe weather forecasts were least successful during fall, with winter the second least successful season. The success of summer forecast relative to other seasons was due to both a greater ability to determine where lightning and severe weather would occur (higher POD score) and a greater ability to determine where it would not occur (higher FOH score). These higher summer scores are possibly a result of the greater coverage of severe weather events that occurs in Europe during the summer months which is caused by larger-scale, easier to forecast weather patterns.

Lightning forecasts did not show consistent improvement; POD and FOH scores increased between year one and year two, but fell back somewhat between year two and year three. Severe weather forecasts improved between each of the three years. However, a data set of three years is not large enough to conclusively determine whether this improvement in forecast score was due to genuine forecast improvement, or whether random variation caused all or most of the improvement.

Similarly, the variation in forecasting success among forecasters was not of a great enough magnitude to conclusively determine whether any one forecaster was consistently any better than any other forecaster. The variation in scores among forecasters may have been due to differences in skill, but other factors such as random variation and the fact that some forecasters forecasted more during difficult seasons and others forecasted more during easier seasons may have played a role.

The severe weather forecasts were useful for decision-making, as their consistently high area under the ROC curve shows. For the vast majority of the forecasting period the 91 forecast-day average was between 0.8 and 0.9, well above the 0.7 threshold necessary for a forecast to be considered useful for a decision-maker.

A major issue that may have adversely affected severe weather verification was the inferior state of severe weather reporting in Europe when compared to the United States. Because large areas of Europe did not receive any severe weather reports during the entire forecast verification period, these areas had to be

excluded from forecast verification. A better severe weather reporting network in Europe would allow verification of ESTOFEX or any other severe weather forecast in Europe with more confidence.

References

Brooks, Harold, 2004: Tornado-Warning Performance in the Past and Future: A Perspective from Signal Detection Theory. *Bulletin of the American Meteorological Society,* **85,** 837-843

Brooks, Harold, Kay, Michael, and Hart, John. Objective Limits on Forecasting Skill of Rare Events.

Doswell, Charles, Davies-Jones, Robert, and Keller, David, 1990: On Summary Measures of Skill in Rare Event Forecasting Based on Contingency Tables. *Weather and Forecasting*, **5**, 576-585

Dotzek, Nikolai, Groenemeijer, Pieter, Feuerstein, Bernold, and Holzer, Alois, 2009: Overview of ESSL's Severe Convective Storms Research Using the European Severe Weather Database ESWD. *Atmospheric Research,* **93**, 575-586.

Finley, John, 1884: Tornado Predictions. *American Meteorological Journal,* **1,** 85-88.

Mason, I., 1982: A Model For Assessment of Weather Forecasts. *Australian Meteorological Magazine,* **30,** 291-303.

Marzban, Caren, 1998: Scalar Measures of Performance in Rare-Event Situations. *Weather and Forecasting,* **13***,* 753-763.

Murphy, Allan, 1996: The Finley Affair: A Signal Event in the History of Forecast Verification. *Weather and Forecasting,* **11,** 3-20.

Murphy, Allan, 1993: What is a Good Forecast? An Essay on the Nature of Goodness in Weather Forecasting. *Weather and Forecasting,* **8**, 281-293.

Murphy, Allan and Winkler, Robert, 1987: A General Framework for Forecast Verification. *Monthly Weather Review,* **115,** 1330-1338.

Roebber, Paul, 2009: Visualizing Multiple Measures of Forecast Quality. *Weather and Forecasting,* **24,** 601-608**.**

Wilks, Daniel S, 1995: *Statistical Methods in the Atmospheric Sciences.* Academic Press, 464 pp.