EVALUATION OF ESTOFEX FORECASTS

H. E. Brooks[1], P. T. Marsh[2], A. M. Kowaleski[3], P. Groenemeijer[4], T. E. Thompson[2], C. S. Schwartz[2,5], C. M. Shafer[2], A. Kolodziej[2], N. Dahl[2], D. Buckey[2]

[1]*NOAA/National Severe Storms Laboratory, 120 David L. Boren Blvd., Norman, Oklahoma, 73072, USA, harold.brooks@noaa.gov*

[2]*School of Meteorology, University of Oklahoma, 120 David L. Boren Blvd., Norman, Oklahoma, 73072, USA*

[3]*National Weather Center Research Experiences for Undergraduates Program and Davidson College, Davidson, North Carolina, 28035, USA*

[4]*European Severe Storms Laboratory, Münchner Str. 20, 82234 Wessling, Germany*

[5]*Current affiliation: National Center for Atmospheric Research, 3450 Mitchell Lane Boulder, Colorado 80301*

(Dated: 15 September 2009)

I. 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, 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 (Fig. 1.)

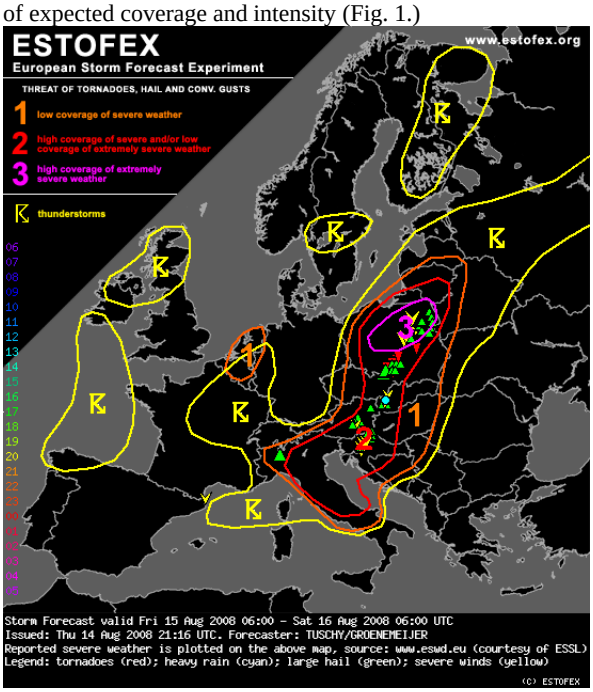

FIG. 1: ESTOFEX forecast issued 14 Aug 2008, valid starting 0600 UTC 15 Aug 2008. Yellow lines indicate regions of expected lightning coverage. Orange, red, and purple lines enclose areas of levels 1, 2, and 3. Observed severe weather reports are shown by symbols.

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.

II. FORECAST AND OBSERVATIONAL DATA

Forecasts are typically issued once per day, usually in the evening, and are valid for a 24-hour period beginning the next morning at 0600 UTC. (Since the forecasters work on a volunteer basis, occasionally forecasts are not issued.) On relatively rare occurrences, updates are issued later. For our purposes, we will consider only the first forecast issued for the day, in order to limit the impacts of additional information being available for the forecasters. We have evaluated three years of forecasts, starting 30 April 2006.

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 on to 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.

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 0.5x0.5 degree 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 EUCLID. The format and area of coverage is somewhat different. The spatial grid is 0.25x0.25 degrees, but the temporal resolution is one hour 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 (0.5x0.5 degrees, one-hour) using the EUCLID domain (Fig. 2). One or more flashes during the 24-hour 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). 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 can determine, a priori, 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.

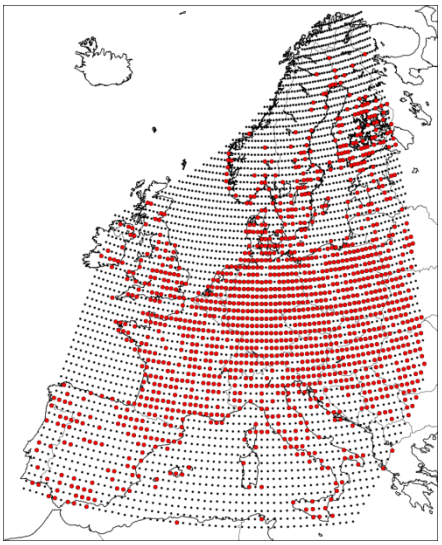

FIG. 2: Verification locations for forecasts. Black dots represent lightning verification locations. Red dots are those locations where severe thunderstorms were reported at least once during the verification period.

III. RESULTS AND CONCLUSIONS

Roebber (2009) introduced a graphical display that is useful for visualizing the performance of dichotomous forecasts of dichotomous events. As such, it is a natural choice for looking at ESTOFEX's lightning forecasts. Plotting the probability of detection (fraction of "yes" events correctly forecasts) versus the frequency of hits (fraction of correct "yes" forecasts) is ideal for considering changes in forecast performance over time (Fig. 3). By computing those quantities over periods of 91 consecutive forecasts, we can see something that resembles a seasonal average, but without restricting our attention to traditional seasons. Clearly, there is a strong seasonal signal, with the forecasts being better in the summer than in the winter. There is significant interannual variability. Peak performance is seen in the second year, at least in terms of the critical success index (CSI).

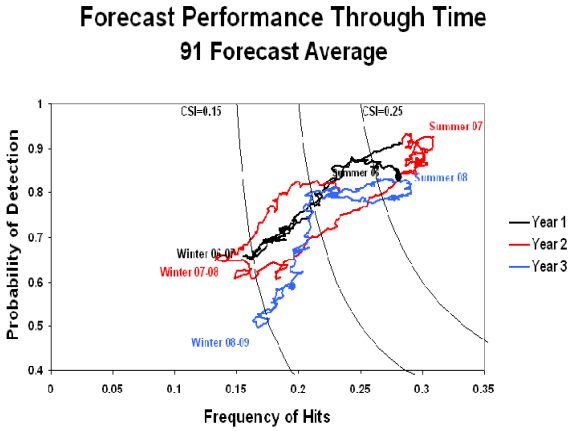

FIG. 3. Running 91-forecast average of probability of detection and frequency of hits for lightning forecasts. Thin black curves represent constant critical success index (CSI=correct forecasts of "yes" events divided by sum of correct forecasts of "yes" events, false alarms, and missed events). Colored lines represent different years of the 3-year evaluation period (red-first year, blue-second year, black-third year.) Perfect forecasts would be located at (1,1).

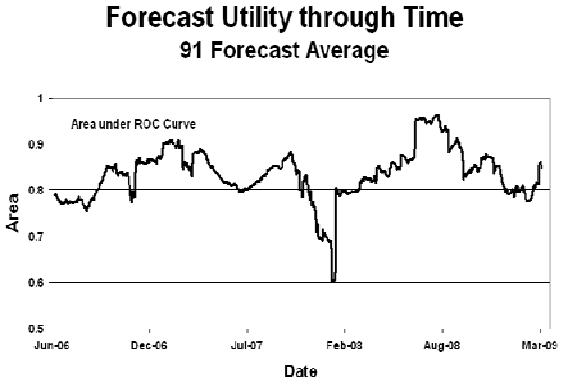

FIG. 4. Area under the curve (AUC) for ROCs as a function of date for 91-forecast running average of severe thunderstorms. Date represents center of the 91-forecast average.

Mason (1982) brought the Relative Operating Characteristics (ROC) curve to the meteorological community. 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 2x2 contingency table from it, and then plotting the probability of detection versus the probability of false detection (fraction of "no" forecasts that have "yes" events). 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.

Again, calculating values from a set of 91 consecutive forecasts is useful for seeing the long-term changes in forecast performance. In contrast to the lightning forecasts, in general there is a long-term increase in forecast performance, but the seasonal signal is not very consistent (Fig. 4). 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 forecast of many events becoming a part of the averaging window following a quiet period of a couple of months.

The ESTOFEX forecasts are of a reasonably good quality and there is evidence (for which this preprint is too short) that differences in forecaster performance are on the order of, or smaller than variability in forecast difficulty.

IV. 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.

V. REFERENCES

Dotzek, N., P. Groenemeijer, B. Feuerstein , and A. Holzer, 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. *Austral. Meteor. Mag.*, **30**, 291-303.

Roebber, P. J., 2009: Visualizing multiple measures of forecast quality. *Wea. Forecasting*, **24**, 601-608.