Assessing the operational skill of predictions of forecast errors

P. J. A. Mailier
Royal Meteorological Institute of Belgium, avenue circulaire 3, 1180 Brussels
Email: pascal.mailier@oma.be Tel.: +32 2 7903966

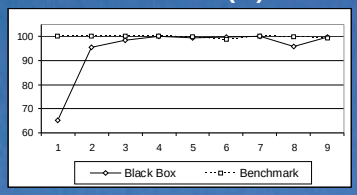
The issue

For energy companies, the ability to guess the most likely direction and magnitude of errors made in deterministic surface temperature forecasts would provide a substantial competitive advantage. An automated, but expensive proprietary system (black box) designed for this purpose has been tested on a set of daily minimum temperature forecasts out to 9 days at one single location in the UK (London Heathrow).

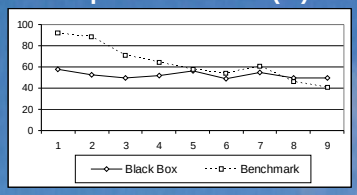
Verification strategy

- A simple and cheap alternative system based on the ECMWF Ensemble Prediction System was devised to serve as benchmark.
- For each of the 9 consecutive forecast days, the two competing systems tried to predict one of three categories:
  - **Blue**: the observed minimum temperature will be *more than 2°C lower than predicted*;
  - **Red**: the observed minimum temperature will be *more than 2°C higher than predicted*;
  - **White**: the observed minimum temperature will be *within 2°C of the predicted value*.
  - **Purple** (= **Blue or Red,** i.e. **not white**) forecasts were treated taking a fuzzy-logic approach: Purple = 50% Blue + 50% Red + 0% White.
- A sample of *76 cases* was collected during the trial, which lasted for nearly 11 weeks (from 20/08/2006 until 03/11/2006).
- 9 daily 3X3 contingency tables were compiled for each contender.
- Performance was assessed quantitatively using several metrics:
  - Association between observed and forecast categories was measured using the χ^2 test for randomness and calculating *1 – p-value* of the test statistic. Some cells in the contingency table have very low counts, so this statistic should be taken as rough guidance only!
  - Proportion of correct forecasts.
  - Hit rate (*H*), false-alarm rate (*F*) and ROC skill score (*ROCSS*). It can be shown that in this case *ROCSS* = *H* – *F*, i.e. the ROC skill score is the same as the maximum attainable forecast value (Jolliffe and Stephenson, 2003).
  - Frequency bias.

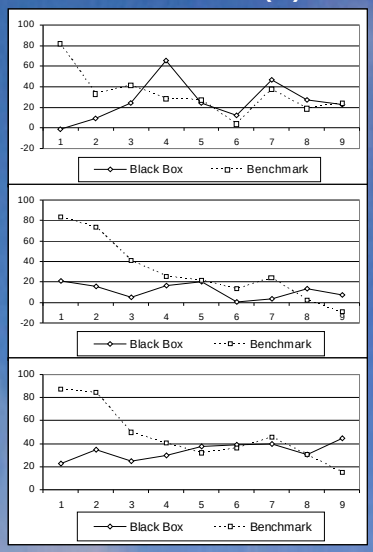
Association (%)



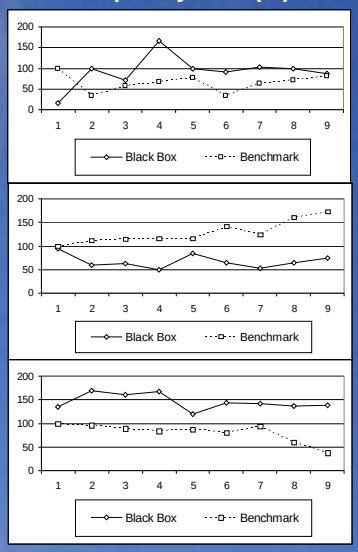
Proportion Correct (%)



ROC Skill Score (%)



Frequency Bias (%)



Blue

White

Red

Summary of results

- Association often close to 100% indicates some forecast skill for both contenders.
- The black box appears less skilful than the benchmark throughout the most predictable range of the forecasts: higher proportion of correct forecasts and higher ROCSS. No evidence of either contender being better than the other in the latter part of the forecast.
- The black box overforecasts Red and underforecasts White whereas the benchmark tends to underforecast Blue.
- Red is easier to predict correctly than White or Blue due to the presence of a significant and systematic cold bias in the minimum temperature forecasts.

Conclusion

The same tests conducted on the corresponding daily maximum temperature forecasts yielded similar results (not shown). No convincing evidence of the superiority of the black box was found and therefore its purchase was not recommended.

Reference

Jolliffe I.T. and Stephenson,D.B. (Eds.), 2003. *Forecast Verification: A Practitioner's Guide in Atmospheric Science*. John Wiley and Sons, 254 pp.