Probabilistic forecasting at Météo-France

Frédéric Atger[1], Météo-France

1. From deterministic to probabilistic forecasts

Until May 1998, end forecasts issued by Météo-France were strictly deterministic. Forecasters now indicate a five level confidence index for every medium-range forecast, from Day+4 to Day+7 (i.e. from +108h to +204h). This confidence index is based on a classification of ensemble forecasts from the Ensemble Prediction System (EPS) of ECMWF. The method of classification is the tubing (Atger 1999). Products of the classification are available through ECMWF dissemination to Member States. Forecasters evaluate the confidence index from an interpretation of these products based on verification statistics and experience (Guyon 1998).

A confidence index can be seen as a very simple probabilistic statement. Confidence indices are probabilistic statements of the first kind, as any statement that indicates in a way or another the expected performance of a deterministic forecast. Statements that indicate explicitly the forecast uncertainty, for instance the expected interval of a quantitative variable or the list of possible instances of a qualitative variable, are probabilistic statements of the second kind. Probabilistic statements of the third kind are probabilities, i.e. quantitative estimates of the likelihood of a weather event.

As pointed out by Murphy (1993), the advantage of probabilistic statements, over deterministic ones, is that they allow forecasters to express more accurately their *judgements*, i.e. their opinions concerning the weather in the future. Human opinions are never completely deterministic, so that a deterministic statement only gives a *truncated* view of a forecaster judgement. End users might still prefer the deterministic form. Deterministic forecasts being easily inferred from probabilistic statements, weather forecasters should express themselves in a probabilistic form in order to be able to satisfy any potential user (Murphy 1993).

Météo-France probabilistic forecasts for the end user have only been so far of the first kind (medium-range forecast confidence index). Medium-range technical guidance, i.e. from the central forecasting office to regional offices, often includes probabilistic statements of the second kind, based on interpretation of EPS products. The aim is now to develop probabilistic statements of the third kind, i.e. probabilities. Probabilities of significant weather events (based on the EPS) start to appear in medium-range technical guidance. Selected professional users (e.g. energy providers) are provided with tailor-made probabilities based on the EPS. The next step will be to provide all end users with probabilities, including televisions and radios.

2. Performance of probabilistic forecasts

Forecast users, as well as weather forecasters, often consider the performance of probabilistic forecasts as a subject of concern. A naive belief is that "probabilistic forecasts are never wrong", which is partly true. Nevertheless, probabilistic forecasts might be more or less performant. Robust

[1]Météo-France (SCEM/PREVI) 42, av. G. Coriolis, 31057 Toulouse cedex
E-mail: frederic.atger@meteo.fr

methods of verification of probabilistic statements have been developed for long, since the advent of probabilistic weather forecasts in the early XXe century (Cooke 1906). The simplest approach is given by the Brier Score (Brier 1950) that is the mean square error of forecast probabilities. **Missing picture**where p_i is the forecast probability of an event, $o_i=1$ if the event occurs, $o_i=0$ if it does not, n the number of forecasts.

The Brier Score is negatively oriented, as any mean square error. The minimum score (0) is only obtained with never wrong deterministic forecasts. Murphy (1973) proposed a widely used decomposition of the Brier Score, analogous to the classical decomposition of the RMSE (Root Mean Square Error), that helps a lot to understand the different aspects of the performance of a probabilistic forecast.

Missing picturewhere the n forecasts have been stratified in m categories grouping the n_k forecasts of probability p_k. The overbar indicates an algebraic mean.

The first term of the decomposition is generally known as *reliability*, or statistical consistency (Talagrand et al. 1997). It expresses the ability to forecast reliable probabilities, i.e. probabilities matching the observed frequencies. In other words, to get a perfect reliability, the event should occur p times out of 100 forecasts of p% probability. An example of (almost) perfect reliability is given by a climatological forecast (unless the season is very anomalous).

The second term of the decomposition is the *resolution*. For a given reliability, the resolution expresses the ability of the forecast probability to vary from day to day. A climatological forecast has no resolution, since the forecast probability is always equal to the climatological frequency. A deterministic forecast has a perfect resolution if never wrong.

The maximum value of the resolution is given by the last term of the decomposition, often called uncertainty, which does not depend on the forecasting system. Uncertainty expresses the intrinsic difficulty to forecast the occurence of an event, i.e. the variance of observations. The uncertainty is maximum (0.25) when the observed frequency is 50%.

3. Calibration of forecast probabilities

Reliability is similar to a bias. For instance, if the observed frequency is 60% when the forecast probability is 80%, the lack of reliability might be expressed as a +20% conditional bias. As any bias, reliability might be improved (i.e. reduced) by statistical correction. In the case of probabilities this correction is known as calibration. When using ensemble prediction, a forecast probability is generally computed as the proportion of ensemble members forecasting a weather event, assuming ensemble members are equally likely to verify. Calibration takes into account the (past) observed frequency of a weather event, given a proportion of ensemble members forecasting this event, to compute the actual probability.

A simple, efficient method of calibration consists in forecasting the (past) observed frequency of an event, given the (actual) proportion of ensemble members forecasting this event (Zhu 1996). The question to answer in this case is: "What is the probability of a weather event, given the forecast?". This *probability calibration* has to be applied separately for every weather event. In case of a quantitative forecast with many significant thresholds (e.g. minimum temperature threshold depending on the user concern), the calibration process has to be repeated a large number of times. Also, some users require a complete probability density function (e.g. to feed specific models) rather than the probability of a given event.

An alternative method, for quantitative forecasts, consists in computing forecast probabilities (for any threshold) from the complete (past) observed distribution of the variable, given an (actual) ensemble distribution. This *distribution calibration* answers the question: "What is the expected distribution of observations, given the forecast?". This second method, initially proposed by Hamill and Colucci (1998) and described in some detail in the next section, is currently under experiment at Météo-France to get reliable forecast probabilities from ECMWF ensemble forecasts.

4. Calibration from Talagrand diagrams

A *Talagrand diagram*, also called rank diagram, shows the distribution of observations given a number of ensemble distributions. It is obtained by counting the number of times the verification is found in each interval between the sorted N ensemble forecasts. Perfect reliability means that the observed frequency in one interval is exactly 1/(N+1), this being true for any interval, whatever its rank. In other words, reliable forecasts exhibit *flat* rank diagrams. In practice, generally U-shaped rank diagrams reveal a lack of ensemble spread: the verification is (relatively) more often found in the tails of the ensemble distribution than in its centre. Calibration based on rank diagrams consists in a transformation of the (actual) ensemble distribution, the probability of each interval being changed from 1/(N+1) to the (past) observed frequency in the corresponding interval.

The main difficulty of the method (as any calibration method) comes from the fact that rank diagrams are built from ensemble distributions that are generally not homogenous. First, an ensemble distribution is representative of a specific location in space and time (a place, a season, a forecast range). Moreover, statistical moments (mean, variance, etc.) are different from an ensemble distribution to another. Ideally, the rank diagram should be built from *identical* ensemble distributions (i.e. having the same moments) valid at the same location, for the same time range, etc. Practically, because of the limited sample that is available after a few years of operational ensemble forecasting (e.g. from December 1996 for the current high resolution version of the ECMWF EPS), ensemble distributions that differ significantly and/or are representative of different locations in space and time are often used together to build a rank diagram (e.g. Strauss and Lanzinger 1995).

5. Calibration and systematic errors

Talagrand diagrams prove very sensitive to systematic errors of individual forecasts. Given a deterministic forecast, model biases result in shifting the mean of the distribution of observations from the forecast value. In the case of an ensemble, a bias results in a L-shaped or a dissymmetric U-shaped rank diagram, the verification being more often "missed" on one side of the distribution (Palany et al. 1999). Systematic errors are rarely purely systematic. They generally depend on the location in space and time, as well as on the forecast value. Therefore systematic errors are likely to compensate one another when building a rank diagram from heterogenous ensemble distributions.

Calibration of forecast probabilities based on such a rank diagram may very well lead to a good *overall* reliability, when verifying an heterogenous distribution of forecast probabilities with the corresponding heterogenous distribution of observations. On the other hand, verification of an homogenous distribution of forecasts (e.g. local forecasts) is likely to reveal a poor reliability. In this respect, calibration resembles any bias correction scheme: the main problem is to base the correction on significant statistics, i.e. to build a sufficiently large homogenous sample. In the case of calibration of forecast probabilities from an ensemble, a solution to get a more homogenous sample is to build the rank diagram from an unbiased ensemble distribution, i.e. after applying a statistical correction to ensemble forecasts.

Two methods are used at Météo-France to reduce local forecast errors of weather parameters. These methods are currently applied to 2m-temperature, 10m-wind and total cloud cover. First, model output statistics (MOS) are applied at every location to individual ensemble forecasts. All forecasts being integrations of the same model, the regression equations have the advantage of being identical for a given location. Then, a first order Kalman filter is used to get rid of recurrent biases. Calibration is only applied to the ensemble distribution once this two steps statistical correction is completed for all ensemble members.

Preliminary results show that statistical correction leads to a definite improvement of the calibration efficiency, calibrated probabilities being more reliable than those obtained through a simple calibration process.

References

Atger, F., 1999. Tubing: an alternative to clustering for the classification of ensemble forecasts. Weather and Forecasting, 14, 5, 741-757.

Brier, G. W., 1950: Verification of forecasts expressed in terms of probability. *Mon. Wea. Rev.*, **78**, 1-3.

Cooke, W.E., 1906. Weighting forecasts. Monthly Weather Review, 34, 274-275.

Guyon, C., 1998: Day 8 forecast at Meteo-France with the Ensemble Prediction System. Expert Meeting on the EPS, Reading, U.K., European Centre for Medium-range Weather Forecasts.

Hamill, T.M. and S.J. Colucci, 1998: Evaluation of Eta-RSM Ensemble Probabilistic Precipitation Forecasts. *Mon. Wea. Rev.*, **126**, 711-724.

Murphy, A., 1973: A new vector partition of the probability score. *J. Appl. Meteor.*, **12**, 595-600.

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

Palany. P, G. Richard, R. Verret, L. Lefaivre, G. Pellerin and P. Houtekamer, 1999. Ten day Temperature anomaly forecast based on ensemble prediction system. Proceedings of the 17th A.M.S. Conference on Weather and Forecasting, Denver, Co, 13-17 Sept. 1999.

Strauss, B. and A. Lanzinger, 1995: Validation of the ECMWF Ensemble Prediction System. Seminar on Predictability, Reading, U.K., European Centre for Medium-range Weather Forecasts, **vol. 2**, 157-166.

Talagrand, O., R. Vautard and B. Strauss, 1997: Evaluation of probabilistic prediction systems. Seminar on Predictability, October 1997, Reading, U.K., European Centre for Medium-range Weather Forecasts.

Zhu, Y., G. Yyengar, Z. Toth, S. M. Tracton and T. Marchok, 1996: Objective evaluation of the NCEP global ensemble forecasting system. *Preprints of the 15th AMS Conference on Weather Analysis and Forecasting*, Norfolk, Virginia, J79-J82.