


Project no. GOCE-CT-2003-505539

Project acronym: ENSEMBLES

Project title: ENSEMBLE-based Predictions of Climate Changes and their Impacts

Instrument: Integrated Project

Thematic Priority: Global Change and Ecosystems

# Deliverable 6.11: Report on the recommendation of methods to evaluate hindcasts probabilistically

Due date of deliverable: May 31, 2007
Actual submission date: Jul 20, 2007

Start date of project: 1 September 2004 Duration: 60 Months

Organisation name of lead contractor for this deliverable LSE

Revision [final]

| **Project co-funded by the European Commission within the Sixth Framework Programme (2002-2006)** | | |
|---|---|---|
| **Dissemination Level** | | |
| **PU** | Public | x |
| **PP** | Restricted to other programme participants (including the Commission Services) | |
| **RE** | Restricted to a group specified by the consortium (including the Commission Services) | |
| **CO** | Confidential, only for members of the Consortium (including the Commission Services) | |

# D6.11 Report on the recommendations of methods to evaluate hindcasts probabilistically

Centre For The Analysis Of Time Series

July 20, 2007

## 1 Introduction

Probabilistic forecasts enable forecasters to express their uncertainty about what they think will happen. This uncertainty information can add great value to a forecast as it enables decision makers to better evaluate their operational risk. Given two distinct probabilistic forecasts, which is better? In order to answer this question we require a method to evaluate probabilistic forecasts along with their uncertainty information. Traditional root mean square measures of distance are ineffective measures of ensemble forecast skill (as they contrast only a point forecast with the verification) and alter-

native evaluation techniques are required to deal with information regarding forecast uncertainty. This work provides an overview of methods to evaluate probabilistic forecasts and their practical application.

## 1.1 Probabilistic Forecasts

A probabilistic forecast is a mathematical object that assigns non-negative forecast weight to all possible events $X$ in the event space $\Omega$. Moreover, the forecast weight is normalised so that the total weight assigned over all possible events sums to 1. Probabilistic forecasts can be either discrete or continuous.

We define a discrete probabilistic forecast, according to the usual axioms of probability, to be a probability measure $P$ over a set of disjoint events $X_i$

$$\Omega \;=\; \bigcup X_i \quad i = 0, \ldots, N-1, \tag{1}$$

so that the intersection between any two distinct events is empty, i.e.

$$X_i \bigcap X_j \;=\; \emptyset \quad i \neq j \tag{2}$$

so that $P(X_i) \geq 0, \forall i$ and $P(\Omega) = 1$. Discrete event spaces with only two events are called dichotomous or binary events, event spaces with more than two possible outcomes are called often multi-category events.

We define a continuous probability forecast of a univariate variable by the probability density function $p(x)$ over the event space $\Omega$. The event space of a univariate variable can usefully be thought of as an interval of the real numbers $\mathbb{R}$. The forecast probability of an event $X$ is then defined by

$$P(X) = \int_{X \in \Omega} p(x)dx \tag{3}$$

where

$$\begin{aligned} P(\Omega) &= \int_{\Omega} p(x)dx \\ &= 1 \end{aligned} \tag{4}$$

These probabilistic forecasts can be formed in any imaginable way. Climatological probabilistic forecasts, based on historical data, are one example, as are probabilistic forecasts based on ensemble forecasts. There are many ways in which one can interpret these different sources of forecast information into a probabilistic forecast. Examples include: best member dressing [22], plotting positions [13], Rank Histogram Recallibration [12], Multiple implementation of single-integration MOS equations [8], Bayesian model averaging [21] and Kernel Dressing [2].

## 1.2 Skill Scores

Forecast evaluation has traditionally relied on the notion of distance, where the quality of a forecast is determined by how close it is to the verification (the quantity that we want to forecast). While the notion of distance is intuitive when comparing point forecasts to point verifications, the conceptual framework does not transfer to evaluating distributions with point verifications as is required when evaluating probabilistic forecasts. This conceptual problem has often been circumvented by casting the probabilistic forecast into a point forecast - most often by computing the distance between the ensemble mean to the verification. While this approach is appealing in its ease of use, reducing a probabilistic forecast to its mean value destroys *all* the uncertainty information and renders fruitless all the effort required to generate a probabilistic forecast.

In order to evaluate full probabilistic forecasts one employs a *skill score*—a function that assigns a scalar value to probability forecast-verification pairs—$(P(x), X)$, where $P(x)$ denotes the forecast probability assigned to all possible outcomes $x$ by either a discrete probability measure or derived from a continuous probability density forecast and $X$ is the verification. There are a multitude of skill scores that one can employ and each has its distinguishing features. We discuss the most relevant properties in the following sections.

A skill score is a function that assigns a scalar value to the forecast-verification pair $(P(x), X)$. We write

$$s = S(P(x), X), \tag{5}$$

to denote the skill $s$ of a probabilistic forecast $P(x)$ given the verification $X$. It is usual to define skill scores in a cost function context so that smaller $s$ is more desirable.

In the same way that there are many metrics with which one can evaluate point forecasts there are many skill scores. The important properties of a skill score are propriety and locality. We discuss these properties in the following.

### 1.2.1 Propriety

Consistency between a forecaster's beliefs and the forecasts issued is referred to as Type 1 goodness by Murphy [20]. Forecast consistency can, however, be detrimentally manipulated by the evaluation methodology. In order to be consistent one should employ a *proper* skill score [14, 24, 10, 3].

Propriety is perhaps best exemplified by a skill score that is not proper. Consider, for example, the Naïve Linear Score defined to be

$$\begin{aligned} s &= S(P(x), X) \\ &= -P(X) \end{aligned} \tag{6}$$

which is simply the negative of the probability assigned to the verification[1]. At first glance this skill score is appealing. It is simple to compute and it is large and negative when a large probabilistic weight is assigned to the verification. The score is, however, not proper, as noted by Staël Von Holstein [14]. We can show that Equation 6 is not proper by showing that we would expect to do better by issuing something other than our beliefs.

Let $p(x)$ be the continuous forecast density that represents the forecaster's beliefs. The expected skill under $p$ is

$$E(S(p(x), x)) \quad = \quad \int -p(z)p(z)dz. \tag{7}$$

Note that $p(x)$ is either a constant, i.e. the forecaster has no preference for any possible outcome, or that there is one event, $\mathbb{X}$ say, that scores better than the expected skill, i.e.

$$-p(\mathbb{X}) < \int -p(z)p(z)dz. \tag{8}$$

Consider the alternative probability density forecast defined by

$$q_\sigma(x) \quad = \quad \frac{1}{\sigma} g\left(\frac{x - \mathbb{X}}{\sigma}\right) \tag{9}$$

defined by an arbitrary smooth, symmetric and normalised kernel $g$, e.g. the Normal distribution, centred at $\mathbb{X}$ with spread $\sigma$. The expected score of this

[1]Note that for continuous probability forecasts we can substitute the density $p(X)$ for the probability forecast $P(X)$ and the results carry through

forecast in the limit as $\sigma \to 0$, given the belief $p$ of the forecaster is then

$$\lim_{\sigma \to 0} \int -q_\sigma(x) p(x) dx \quad = \quad -p(\mathbb{X}). \tag{10}$$

The expected score of the forecast $q_\sigma$ under $p$ is better than the expected score of $p$ and any reasonable forecaster being evaluated using the Naive Linear Score would issue $q_\sigma$ rather than $p$. The Naïve linear Score thus encourages the forecaster to issue forecasts that are different to their beliefs and hence the skill score is improper.

A skill score is proper if for any two probability densities $p(x)$ and $q(x)$

$$\int S(q(x), z) p(z) dz \geq \int S(p(x), z) p(z) dz. \tag{11}$$

A score is strictly proper if this inequality is achieved only if $p(x) = q(x)$ for all $x$.

### 1.2.2 Locality

Another property of skill scores that is of interest is locality. A skill score is local if it depends only on the probability assigned to the verification, i.e

$$S(P(x), X) \quad = \quad S(P(X), X). \tag{12}$$

The quality of a forecast, as measured by a local score, depends only on the probability assigned to the verification and not on the shape of the distribu-

tion. This property has been referred to as *relevance* by Staël Von Holstein [14].

Local scores are intuitively appealing, rewarding those forecasts that assign large probabilities to those events that occur. There are, however, a number of nonlocal scores that are widely used. What is the need for nonlocal skill scores? Murphy [17] suggests occasions when the 'distance' of the forecast from the verification is important, for example in cases of ordered variables such as temperature and precipitation. For those cases, Murphy argues, one wants to reward those distributions that assign probability near the verified outcome. In order to motivate the usefulness of a concept of distance Murphy puts forward the following example:

> ..., consider the forecasts
>
> $$\begin{aligned} \mathbf{r} &= (0, 0.1, 0.3, 0.4, 0.2) \\ \mathbf{r}' &= (0, 0.3, 0.1, 0.4, 0.2) \end{aligned}$$
>
> on an occasion when [the fourth class, $p(x_4) = 0.4$] occurs. Note that $\mathbf{r}$ and $\mathbf{r}'$ consist of the same probabilities and that $\mathbf{r}_4 = \mathbf{r}'_4 = 0.4$; the difference between $\mathbf{r}$ and $\mathbf{r}$' is simply that $\mathbf{r}_2$=$\mathbf{r}'_3$ and $\mathbf{r}_3 = \mathbf{r}'_2$. The PS [Brier score] would assign the $\mathbf{r}$ and $\mathbf{r}'$ the same score (0.5). However, if the variable of concern is ordered, many

> meteorologists and others would consider, no doubt, **r** to be a better forecast than **r**'

Although one might prefer forecast **r** to **r**$'$ it is important to note that we are not evaluating the quality of a single forecast but rather the quality of a forecast system. In order to do this we must base our evaluation on a series of forecasts.

There are proper local scores and proper non-local scores. Is there any reason why one should be chosen over the other? At present, there is no compelling argument for choosing a skill score based on considerations of locality.

A separate motivation for nonlocal skill scores is the need to deal with verifications that are assigned zero probability. Again, these deficiencies in the forecast can be addressed by the skill score, or rather one can ensure that no event is assigned zero probability.

# 2 Skill Scores

This section reviews a number of proper skill scores.

## 2.1 Brier Skill Score

The Brier skill score [1] for continuous forecasts is defined to be

$$\begin{aligned} S(p(x), X) &= \int (p(z) - \delta(X - x))^2 dz \\ &= 1 - 2p(X) + \int p(z)^2 dz, \end{aligned} \tag{13}$$

where $\delta(y)$ is the delta function: $\delta(y) = 0, y \neq 0$ and infinity at $y = 0$. For discrete forecasts, the Brier score is

$$S(P(x), X) = \sum_{i}^{N} (P(X_i) - \delta(X_i - X)))^2 \tag{14}$$

The Brier score is frequently applied to forecasts of dichotomous events, e.g. rain/no rain. In such cases the events map onto the binary events $X_0 = 0$ (no rain) and $X_1 = 1$ (rain) and the skill score can be reduced to

$$S(p(x), X) = 2 \cdot (x - p(X = 1))^2 \tag{15}$$

To show that the Brier score is proper we consider the expected skill of the forecast $p(x)$.

$$\begin{aligned} E(S(p(x), x)) &= \int p(y)(1 - 2p(y)) + \int p(z)^2 dzdy \\ &= 1 - 2 \int p(z)^2 dz + \int \int p(y)p(z)^2 dzdy. \end{aligned} \tag{16}$$

The expected skill score for a different forecast $q(x)$, given the belief $p(x)$ is

$$\begin{aligned} E(S(q(x), x)) &= \int p(y)(1 - 2q(x)) + \int q(z)^2 dzdy \\ &= 1 - 2 \int p(z)q(z)dz + \int \int p(y)q(z)^2 dzdy. \end{aligned} \tag{17}$$

If the Brier skill score is improper then we can improve our score by issuing something other than our beliefs, i.e. we can have

$$\begin{aligned} E(S(p(x),x) &\geq E(S(q(x),x)) \\ \int\int p(y)p(z)^2 dzdy - 2\int p(z)^2 dz &\geq \int\int p(y)q(z)^2 dzdy - 2\int p(z)q(z)dz \\ 0 &\geq \int p(z)^2 dz + \int q(z)^2 dz - 2\int p(z)q(z)dz \\ 0 &\geq \int (p(z)-q(z))^2 dz. \end{aligned} \tag{18}$$

Clearly, this cannot be the case and hence the Brier skill score is proper.

The Brier score is not local for continuous forecasts and multi-category discrete forecast, since the score depends on the probabilities of events forecast that do not occur. The score is, however, straightforward to implement in the case of discrete forecasts.

### 2.1.1 Log Probability Score

The log probability score [11], or *ignorance* skill score [23], is defined to be

$$S(p(X),x) = -\log(p(x)). \tag{19}$$

The skill score is proper [10, 15] and local, since $S$ depends only on the forecast at $x$. It is the only proper local skill score for continuous forecast distributions. Propriety can be shown by considering the expected skill of two forecasts $p(x)$ and $q(x)$ given the belief $p(x)$.

$$
\begin{aligned}
E[S(p(X),x)] - E[S(q(X),x)] &= \int -p(x)\log p(x) - \int -p(x)\log q(x)dx \\
&= \int p(x)\log\frac{q(x)}{p(x)}dx. \qquad (20)
\end{aligned}
$$

The term on the right is zero if and only if $p(x) = q(x)$ and is positive otherwise. One cannot expect to improve one's score by issuing anything other than one's beliefs $p(x)$.

The log probability score is straightforward and computationally economical to implement for both discrete and continuous forecast distributions. When comparing two forecast systems it is important that they are computed over the same event space, i.e. to compare a discrete forecast and a continuous density function one must compute the probability measure assigned by the continuous forecast to the event space implicit in the discrete forecast.

There have been criticisms of the log probability score, most notably on its scoring of forecasts that assign zero probability to the outcome. The property has been defined by Selten [26] as hypersensitivity. Whether or not hypersensitivity is a downside depends largely on whether or not the forecaster is willing to accept zero probability being ascribed to events that actually occur.

### 2.1.2 Ranked Probability Score

The ranked probability score [7, 15] is a scoring rule, based on the quadratic scoring rule, that seeks to reward forecasters for assigning probability, not only to the verified event, but also to near the verified event. The continuous ranked probability score is defined to be

$$S(p(x), X) = \int (f(x) - H(x - X))^2 dx, \tag{21}$$

where $f(x)$ is the cumulative density function of $p(x)$ defined by

$$f(x) = \int_{-\infty}^{x} p(y) dy \tag{22}$$

and $H(y)$ is the Heaviside function, which is zero for $y < 0$ and 1 for $y \geq 0$. The score corresponds to the integral of the Brier score for the associated binary probabilistic forecasts at all real value thresholds.

For discrete forecasts the ranked probability score is defined by

$$S(P(x), X) = \sum_{i=0}^{N-1} (F(X_i) - H(X))^2, \tag{23}$$

where $F(x)$ is the cumulative probability function of $P(x)$. The ranked probability score is proper [10, 16] but not local.

The continuous ranked probability score has a convenient representation

when the predictive distribution is normal, $\mathbb{N}(\mu, \sigma)$ [10], with

$$S(\mathbb{N}(\mu,\sigma),X) = \frac{\sigma}{\sqrt{\pi}}\left(1-\sqrt{\pi}\left(\frac{X-\mu}{\sigma}\right)\mathrm{erf}\left(\frac{X-\mu}{\sqrt{2\sigma^2}}\right) -\sqrt{2}\exp\left(-\frac{(X-\mu)^2}{2\sigma^2}\right)\right) \quad (24)$$

where erf denotes the error function (or Gauss error function). This non-local score is however, more difficult to evaluate for non-standard predictive distributions, such as those derived by kernel dressing each ensemble member.

# 3 Evaluation By Skill Score

The performance of a probabilistic forecasting system can only be evaluated over a series of forecasts. That is one requires a set of forecast-verification pairs $\{p_t(X), x_t\}, t = 1, \ldots, N$ in order to determine the quality of a forecasting system. It is usual to summarise the performance of a forecasting system by the sample skill score,

$$\bar{s} = \frac{1}{N}\sum_{t=1}^{N} S(p_t(X), x_t). \quad (25)$$

The sample skill score $\bar{s}$ is an estimate of the "true skill" $s$. It is important to estimate the uncertainty in this estimate. If we believe that the uncertainty in the sample statistic is Gaussian then we can compute the standard error. An alternative expression for the uncertainty can be achieved by bootstrapping

[5] the sample skill score. Bootstrapping involves drawing, with replacement, from the sample of skill scores and computing the sample mean of these bootstrapped samples. The result is a distribution of sample means that expresses our uncertainty in the estimate of the sample mean.

One often wants to compare the performance of two models, say $p(X)$ and $q(X)$. A straightforward comparison of the sample means and their uncertainties can be misleading. It may be the case, for example, that the sample means are statistically indistinguishable, i.e. the uncertainty in the sample means overlap at some significant level. One is often concerned with day-to-day performance and a comparison of the mean performance does not necessarily provide this information. To investigate the day-to-day performance one can consider the sample mean of the relative skill scores

$$\bar{D} \;=\; \frac{1}{N}\sum_{t=1}^{N} S(p_t(X), x_t) - S(q_t(X), x_t). \tag{26}$$

Given that smaller skill scores are more desirable, positive $\bar{D}$ indicates that the forecast system that produces $q(X)$ is better, while negative $\bar{D}$ indicates that $p(X)$ is better. Again, one can estimate the uncertainty in $\bar{D}$ using bootstrap error bars and require that these bars are above or below zero in order to find in favour of either forecasting system.

# 4 Summary

Probabilistic forecasts can be evaluated using one of a number of skill scores. It is important that whatever skill score is employed that it does not encourage the forecaster to issue something other than their beliefs. A comparison of forecast systems should take into account the relative performance over each forecast instance. This can be achieved by considering the relative skill score. Moreover, it is important to remember that the skill over a finite set of forecast-verification pairs is itself a statistic with associated uncertainty. A fair evaluation will present this uncertainty information.

# References


[1] G.W. Brier. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1–3, January 1950.

[2] J. Bröcker and L.A. Smith. From ensemble forecasts to predictive distribution functions. submitted to Tellus.

[3] J. Bröcker and L.A. Smith. Scoring probabilistic forecasts: On the importance of being proper. *Weather and Forecasting*, 22(2):382–388, 2007.

[4] M.H. DeGroot and S.E. Fienberg. The comparison and evaluation of forecasters. *The Statistician*, 32:12–22, 1983.

[5] B. Efron and R. Tibshirani. Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical Science*, 1:54–77, 1986.

[6] M. Ehrendorfer and A.H. Murphy. Comparative evaluation of weather forecasting systems: sufficiency, quality, and accuracy. *Monthly Weather Review*, 116:1757–1770, 1988.

[7] E.S. Epstein. A scoring system for probability forecasts of ranked categories. *Journal of Applied Meteorology*, 8:985–987, December 1969.

[8] M.C. Erickson. Medium-range prediction of PoP and max/min in the era of ensemble model output. In *Preprints, 15th Conference on Weather and Forecasting*, pages J35–J38. Am. Meteorol. Soc.

[9] D. Friedman. Effective scoring rules for probabilistic forecasts. *Management Science*, 29(4):447, April 1983.

[10] T. Gneiting and A.E. Raftery. Strictly proper scoring rules, prediction, and estimation. Technical Report 463, Department of Statistics, University of Washington, September 2004.

[11] I.J. Good. Rational decisions. *Journal of the Royal Statistical Society, Ser. B*, 14:107–114, 1952.

[12] T.M. Hamill and S.J. Colucci. Evaluation of Eta-RSM ensemble probabilistic precipitation forecasts. *Mon. Weather Rev.*, 126:711–724, 1998.

[13] H.L. Harter. Another look at plotting positions. *Commun. Stat. A-Theor.*, 13:1613–1633, 1984.

[14] C.-A.S. Staël Von Holstein. Measurement of subjective probability. *Acta Psycologica*, 34:146–159, 1970.

[15] J.E. Matheson and R.L. Winkler. Scoring rules for continuous probability distributions. *Management Science*, 22(10):1087, June 1976.

[16] A.H. Murphy. On the r̈anked probability score*:* *Journal of Applied Meteorology*, 8:988–989, December 1969.

[17] A.H. Murphy. The ranked probability score and the probability score: a comparison. *Monthly Weather Review*, 98(12):917–924, December 1970.

[18] A.H. Murphy. Forecast verification: its complexity and dimensionality. *Monthly weather review*, 119:1590–1601, 1991.

[19] A.H. Murphy. What is a good forecast? an essay on the nature of goodness in weather forecasting. *Weather and Forecasting*, 8:281–293, June 1993.

[20] A.H. Murphy and R.L. Winkler. A general framework for forecast verification. *Monthly Weather Review*, 115:1330–1338, July 1987.

[21] A.E. Raftery, F. Balabdaoui, T. Gneiting, and M. Polakowski. Using Bayesian model averaging to calibrate forecast ensembles. *Mon. Weather Rev.*, 133:1155–1174, 2005.

[22] M. Roulston and L. Smith. Combining dynamical and statistical ensembles. *Tellus A*, 55:16–30, 2003.

[23] M.S. Roulston and L.A. Smith. Evaluating probabilistic forecasts using information theory. *Monthly Weather Review*, 130:1653–1660, 2002.

[24] L.J. Savage. Elicitation of personal probabilities and expectations. *Journal of the American Statistical Association*, 66(336):783–801, December 1971.

[25] M.J. Schervish. A general method for comparing probability assessors. *The Annals of Statistics*, 17(4):1856–1879, 1989.

[26] R. Selten. Axiomatic characterization of the quadratic scoring rule. *Experimental Economics*, 1(43-62), 1998.