Proper scores

- A score for a probabilistic forecast is a summary measure that evaluates the probability distribution. This condenses all the information into a single number and can be potentially misleading.
- Let us assume that we predict the distribution \( p_{fc}(x) \) while the verification is distributed according to a distribution \( p_y(x) \). Not all scores indicate maximum skill for \( p_{fc} = p_y \).
- A score (or scoring rule) is \((strictly)\) proper if the score reaches its optimal value if (and only if) the predicted distribution is equal to the distribution of the verification.
- If a forecaster is judged by a score that is not proper, (s)he is encouraged to issue forecasts that differ from what her/his true belief of the best forecast is! In such a situation, one says that the forecast is hedged or that the forecaster plays the score.
- Examples of proper scores are: Brier Score, continuous (and discrete) ranked probability score, logarithmic score
- see Gneiting and Raftery (2007) for more details
Example of a score that is not proper

- consider the linear score: \( \text{LinS} = |p - o| \)
- dichotomous event \( e \): \( e \) occurred \( (o = 1) \), \( e \) did not occur \( (o = 0) \)
- assume the event occurs with the true probability of 0.4
- If the prediction is 0.4, the \textit{expected} linear score is
  \[
  E(\text{LinS}) = 0.4|0.4 - 1| + (1 - 0.4)|0.4 - 0| = 0.48
  \]
- If the prediction is instead 0, the expected linear score is
  \[
  E(\text{LinS}) = 0.4|0 - 1| + (1 - 0.4)|0 - 0| = 0.40
  \]

Note, that it is easy to prove that the Brier score is strictly proper (e.g. Wilks 2011)

An example with two proper score

Simple idealised example

We compare Alice’s and Bob’s forecasts for \( Y \sim \mathcal{N}(0, 1) \),

\[
F_{\text{Alice}} = \mathcal{N}(0, 1) \quad F_{\text{Bob}} = \mathcal{N}(4, 1)
\]

Based on 10,000 forecast experiments,

<table>
<thead>
<tr>
<th>Forecaster</th>
<th>CRPS</th>
<th>LogS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alice</td>
<td>0.56</td>
<td>1.42</td>
</tr>
<tr>
<td>Bob</td>
<td>3.53</td>
<td>9.36</td>
</tr>
</tbody>
</table>
A conditional sample for evaluating Alice and Bob

Simple toy example

Based on the 10 largest observations,

<table>
<thead>
<tr>
<th>Forecaster</th>
<th>CRPS</th>
<th>LogS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alice</td>
<td>2.70</td>
<td>6.29</td>
</tr>
<tr>
<td>Bob</td>
<td>0.46</td>
<td>1.21</td>
</tr>
</tbody>
</table>

The forecaster’s dilemma

More generally, for non-constant weight functions $w$, any scoring rule

$$S^*(F, y) = w(y)S(F, y)$$

is improper even if $S$ is a proper scoring rule (Gneiting and Ranjan, 2011). Here, $y$ and $F$ denote the verifying observation and the predicted distribution, respectively.

Forecaster’s dilemma

Forecast evaluation only based on a subset of extreme observations corresponds to improper verification methods and is bound to discredit skillful forecasters.

Acknowledgement: Forecaster’s dilemma and Alice and Bob’s forecast based on slides provided by Sebastian Lerch (Heidelberg Institute for Theoretical Studies), see also http://arxiv.org/pdf/1512.09244
Scores for probabilistic/ensemble forecasts of continuous scalar variables

some (but not all) useful measures

- RMSE and other scores used for single forecasts applied to ensemble mean
- rank histograms (reliability again)
- continuous ranked probability score (reliability and resolution)
- logarithmic score (for Gaussian) (reliability and resolution)
- reliability of the ensemble spread (domain-integrated and local)

Continuous ranked probability score

\[
\text{CRPS} = \text{Mean squared error of the cumulative distribution } P_{\text{fc}}
\]

\[
cdf \text{ of observation } \quad P_y(x) = P(y \leq x) = H(x - y) = 1\{y \leq x\}
\]

\[
cdf \text{ of forecast } \quad P_{\text{fc}}(x) = P(x_{\text{fc}} \leq x)
\]

Here, \( H \) and \( 1 \) denote the Heaviside step function and the indicator function, respectively.

\[
\text{CRPS} = \int_{-\infty}^{+\infty} (P_{\text{fc}}(x) - P_y(x))^2 \, dx = \int_{-\infty}^{+\infty} BS_x \, dx
\]

equal to mean absolute error for a single forecast
How to compute the CRPS

The integral \( \int \ldots dx \) can be evaluated exactly by using the intervals defined by the \( M \) ensemble forecasts and the verification rather than some fixed interval \( \Delta x \):

\[
\text{CRPS} = \sum_{j=0}^{M} c_j
\]

\[
c_j = \alpha_j p_j^2 + \beta_j (1 - p_j)^2
\]

\[
p_j = j / M
\]

For a Gaussian distribution an analytical formula for the CRPS is available.

Assume that the predicted Gaussian has mean \( \mu \) and variance \( \sigma^2 \) and that the verification is denoted by \( y \).

\[
\text{CRPS} = \frac{\sigma}{\sqrt{\pi}} \left[ -1 + \sqrt{\pi} \frac{y - \mu}{\sigma} \Phi \left( \frac{y - \mu}{\sqrt{2\sigma}} \right) + \sqrt{2} \exp \left( -\frac{(y - \mu)^2}{2\sigma^2} \right) \right]
\]

Here, \( \Phi \) denotes the error function \( \Phi(x) = \frac{2}{\sqrt{\pi}} \int_0^x \exp(-t^2) \, dt \).

This relationship is particularly useful for calibration purposes (Non-homogeneous Gaussian regression).
The CRPS can be decomposed into a reliability component and a resolution component.

The CRPS is additive: The CRPS for the union of two samples is the weighted (arithmetic) average of the CRPS of the two samples with the weights proportional to the respective sample sizes.

The components of the CRPS are not additive. The components can be computed from the sample averages of the $\alpha_j$ and $\beta_j$ distances.

This is similar to the decomposition of the Brier score. However, the reliability (resolution) component of the CRPS is not the integral of the reliability (resolution) component of the Brier scores.

The reliability component of the CRPS is related to the rank histogram but not identical.

see Hersbach (2000) for details

CRPS with threshold-weighting

Can be used for instance to focus on the tails of the climatological distribution, e.g. strong wind, intense rainfall.

The threshold-weighted CRPS weights the integrand (= Brier score for threshold $z$)

$$\text{twCRPS}(F, y) = \int_{-\infty}^{\infty} (F(z) - \mathbb{1}\{y \leq z\})^2 w(z) dz$$

$w(z)$ is a weight function on the real line. The score twCRPS is proper and avoids the problem with looking only at a sample of extreme outcomes (Alice and Bob’s example).

Gneiting, T. and Ranjan, R. (2011)

Ranked Probability Score (RPS)

- The CRPS $\int BS_x \, dx$ has a discrete analog, the (discrete) ranked probability score:

$$RPS = \sum_{k=1}^{L} BS_{x_k} = \sum_{k=1}^{L} \left( P_{fc}(k) - P_y(k) \right)^2$$

- The thresholds $x_k$ that separate the $L$ categories can be chosen in various ways:
  - equidistant ($RPS \rightarrow CRPS$ as $\Delta x \rightarrow 0$)
  - climatologically equally likely, e.g. tercile boundaries

Logarithmic score

- For a forecast consisting of a probability density $p_{fc}(x)$, define

$$LS = - \log(p_{fc}(y))$$

where $y$ denotes the observation (or analysis).
- This score is proper and local.
- ensemble forecasts $\rightarrow$ probability density
- A simple yet useful exercise is to use the Gaussian density given by the ensemble mean $\mu$ and the ensemble variance $\sigma^2$. Then, the logarithmic score is given by

$$LS = \frac{(\mu - y)^2}{2\sigma^2} + \frac{1}{2} \log(2\pi\sigma^2)$$

- Thus, it consists of the squared error of the ensemble mean normalized by the ensemble variance and a logarithmic term that penalizes large variance. The first term is a measure of the reliability and the second term is a measure of the sharpness of the forecast.
Daily EPS stdev (shaded) and ens. mean (cont.)

500 hPa geopotential (m² s⁻²) at 72 h lead; init. time 6 December 2010

Spread-reliability methodology

consider (local) pairs of ensemble variance and squared error of the ensemble mean — stratified by the ensemble variance
Verification of ensembles and single forecasts

- When monitoring an operational forecasting system that consists of single (unperturbed) forecasts and an ensemble, it is useful to compare changes in the performance of the ensemble with changes seen for the single forecast(s).
- But what scores should be compared when looking at a single forecast versus an ensemble?
- Many scores for ensembles are meaningful when computed for single forecasts
- equivalences
  - CRPS — MAE
  - BS — BS single fc (using probabilities 0 and 1)
- Obviously, probabilistic skill of a “naked” (= raw) single forecast is inferior to the probabilistic skill of a dressed single forecast. The dressing kernel can be estimated from past error statistics.
Dressed control forecast: \( v \ 850 \text{ hPa}, \ 35^\circ - 65^\circ N, \ DJF09 \)

- **EPS**  
  raw prob. for CRPS; Gaussian for LS

\[ N(\text{CF}, \sigma^2_{\text{err}}(\text{CF})) \]

\( \sigma_{\text{err}} \) estimated from reforecasts

**CRPS**

**LS**

- EM more accurate than CF \( \Rightarrow \) this permits a sharper Gaussian distribution.
- The Logarithmic score discriminates better the value of flow-dependent variations in ensemble variance than the CRPS.
Uncertainty of the verifying observations
or, more generally, the verifying data

- In real applications the true state $x_t$ of the atmosphere is not know exactly. The observation $y$ has an error
  \[ y = x_t + \epsilon \]

- Assume an ensemble is perfectly reliable, i.e. ensemble members $x_e \sim \rho_e$ and the true state $x_t \sim \rho_t$ are realisations of the same distribution $\rho_e = \rho_t$.
- Then, the observation $y$ is a realisation of the distribution given by the convolution of the true distribution and the error distribution
  \[ \rho_y = \rho_t * \rho_\epsilon \]
- Thus, a verification with respect to $y$ will indicate a lack of reliability.

Verification in the presence of observation uncertainties

- solution: postprocess ensemble members prior to verification
- verify ensemble members to which noise has been added:
  \[ x_E = x_e + \epsilon \quad \text{with} \quad \epsilon \sim \rho_\epsilon \]
- Then $\rho_E = \rho_y$
The climatological distribution
temperature in 850 hPa
15 March (based on ERA-Interim 1989–2008)

contours: mean — shading: stdev

Fictitious skill due to a poor climatological distribution

• If one uses the same climatological distribution for a domain with different climatological characteristics (mean, stdev, . . . ), the skill with respect to that distribution is not real skill. It reflects the poor quality of the climatological distribution.

• Same applies if seasonal variations of the climatological distribution are not represented.

• This criticism applies for instance if the climatological distribution is derived from the verification sample itself by aggregating different start times and different locations.

• It can also be misleading to compare skill scores from different prediction centres when the skill scores have been computed against own analyses.

• If the same climatological distribution (say ERA-Interim) is used as reference, this climatological distribution has the lowest skill when verified against the analysis that deviates most from the analyses used for computing the climatological distribution.
Comparing model versions/ numerical experiments

- 46 cases (1 year, every 8 days)
- Could difference in score be a result of chance?
- How large does a difference have to be to be trusted?
- Case-to-case variability of predictability implies distribution of score for given lead time is fairly wide
- ⇒ not easy to get enough cases to distinguish score distributions of two numerical experiments

95% confidence intervals

- 46 cases (1 year, every 8 days)
- Variability of score differences is much smaller!
- ⇒ Paired sample of cases (start dates)
- For each forecast lead time, consider sample of score differences

- Temporal auto-correlation taken into account using AR(1) model when estimating variance of mean difference
More verification topics

- sensitivity to ensemble size and estimation of verification statistics in the limit $M \to \infty$
- skill on different spatial scales
- multivariate aspects
- decision making and verification

Probabilities can help with making decisions

- Open air restaurant scenario:
  - to open additional tables costs £20 and provides £100 extra income (if $T > 24^\circ C$)
  - On a particular day, the forecast is $P(T > 24^\circ C) = 0.30$
  - What should the restaurant do?

- Compute the profit/loss (£) over 100 days (assuming reliable probabilities):

  profit on warm days($T > 24^\circ C$) = 30 × (100 − 20) = +2400
  profit on cool days($T \leq 24^\circ C$) = 70 × (0 − 20) = −1400
  total profit = +1000

- It is profitable to open additional tables if the probability of a warm day exceeds 0.20.
- The ratio of cost to loss (or cost to extra profit) determines at what probability value it is beneficial to take action. For low (high) cost/loss, action should be taken already (only) if the event is predicted with a low (high) probability.
## Decision making — cost loss model

<table>
<thead>
<tr>
<th>Action taken</th>
<th>Yes</th>
<th>No</th>
</tr>
</thead>
<tbody>
<tr>
<td>Event occurs</td>
<td>C</td>
<td>L</td>
</tr>
<tr>
<td>Event forecast</td>
<td>a</td>
<td>c</td>
</tr>
</tbody>
</table>

### Potential costs

- Expense when using climatological prob. \( E_c = \min(C, \bar{C}L) \)
- Expense when using a perfect forecast \( E_p = \bar{C}C \)
- Expense when using the forecast \( E_f = aC + bC + cL \)

### Saving

\[ V = \frac{\text{saving from using forecast}}{\text{saving from using perfect fc.}} = \frac{E_c - E_f}{E_c - E_p} = \min(\alpha, \bar{\sigma}) - F(1 - \bar{\sigma})\alpha + H\bar{\sigma}(1 - \alpha) - \bar{\sigma} \]

where

\[ \alpha = \frac{C}{L}; \quad H = \frac{a}{a+c}; \quad F = \frac{b}{b+d}; \quad \bar{\sigma} = a + c \]

---

### (Potential) economic value

**Northern Extra-Tropics (winter 01/02) D+5 FC > 1mm precipitation**

- maximum value reached at \( \alpha = C / L = \bar{\sigma} \)
- maximum value for all \( C / L \) is \( \max V = H - F \)
- when a (reliable) probabilistic forecast predicts an event with probability \( p \), all users with \( C / L < p \) should act.
- one speaks of potential economic value if calibrated probabilities are used to make decision
Decision making — weather roulette

The funding agency of a weather forecast centre believes that the forecasts are useless and not better than climatology!

The Director of the weather centre believes that their forecasts are more worth than a climatological forecast!

She challenges the funding agency in saying: 
*I bet, I can make more money with our forecasts than you can make with a climatological forecast!*

*Hagedorn and Smith (2009)*
Decision making — weather roulette

- Both parties, the funding agency (A) and the Director (D), agree that both of them open a weather roulette casino, and that both of them spend each day 1 k€ of their own budget in the casino of the other party.

- A & D use their favourite forecast to (i) set the odds of their own casino and (ii) distribute their money in the other casino.
  - A sets the odds of its casino and distributes the stake according to climatology.
  - D sets the odds of her casino and distributes her stake according to her forecast.

- They agree to bet on the 2m-temperature at London-Heathrow being well-below, below, normal, above, or well-above the long-term climatological average (5 possible categories).

Decision making — weather roulette

- Odds in casino A: \( o_A(i) = \frac{1}{p_A(i)} \)  
  casino D: \( o_D(i) = \frac{1}{p_D(i)} \)

  with: \( i=1,...,N \): possible outcomes  
  \( p_A(i) \): A’s probability of the \( i \)th outcome  
  \( p_D(i) \): D’s probability of the \( i \)th outcome

- Stakes of A: \( s_A(i) = p_A(i) \times c \)  
  of D: \( s_D(i) = p_D(i) \times c \)

  with: \( c = \) available capital to be distributed every day

- Return for A: \( r_A(v) = o_D(v) \times s_A(v) = \frac{p_A(v)}{p_D(v)} \times c \)

- Return for D: \( r_D(v) = o_A(v) \times s_D(v) = \frac{p_D(v)}{p_A(v)} \times c \)

  with: \( v = \) verifying outcome
Decision making — weather roulette

3-day forecasts

Verification bin
EPS

probability of verification bin

accumulated winnings for EPS

Decision making — weather roulette

10-day forecasts

Verification bin
EPS

probability of verification bin

accumulated winnings for EPS

weather roulette capital gains are closely related to the logarithmic score


Leutbecher, M., 2009: Diagnosis of ensemble forecasting systems. In *Seminar on Diagnosis of Forecasting and Data Assimilation Systems*, ECMWF, Reading, UK, 235–266.


M. Leutbecher Ensemble Verification II