Ensemble verification and extreme events

Chris Ferro

Department of Mathematics
University of Exeter, UK

ECMWF Annual Seminar (Reading, 14 September 2017)

Probability forecasts

Use proper scores to rank probability forecasts

Consider forecasts $f_1, f_2, \ldots$ and outcomes $x_1, x_2, \ldots$

Definition: A **scoring rule**, $s(f, x)$, gives a numerical score to each forecast.

Example: Let $x = 0$ or 1, $f = \Pr(x = 1)$ and $s(f, x) = (f - x)^2$.

Measure performance by the mean score, $\bar{s} = \sum_{i=1}^{n} s(f_i, x_i)/n$.

Use proper scores to rank probability forecasts

Consider forecasts $f_1, f_2, \ldots$ and outcomes $x_1, x_2, \ldots$

Definition: A **scoring rule**, $s(f, x)$, gives a numerical score to each forecast.

Example: Let $x = 0$ or 1, $f = \Pr(x = 1)$ and $s(f, x) = (f - x)^2$.

Measure performance by the mean score, $\bar{s} = \sum_{i=1}^{n} s(f_i, x_i)/n$.

Suppose that $x_1, x_2, \ldots$ have frequency distribution p and that we issue the same forecast, f, for all $x_1, x_2, \ldots$

The best choice is $f = p$.

Definition: A scoring rule is **proper** if the long-run mean score is optimized by $f = p$.

Example: $(f - x)^2$ is proper; $|f - x|$ is improper.

Example: two proper scores

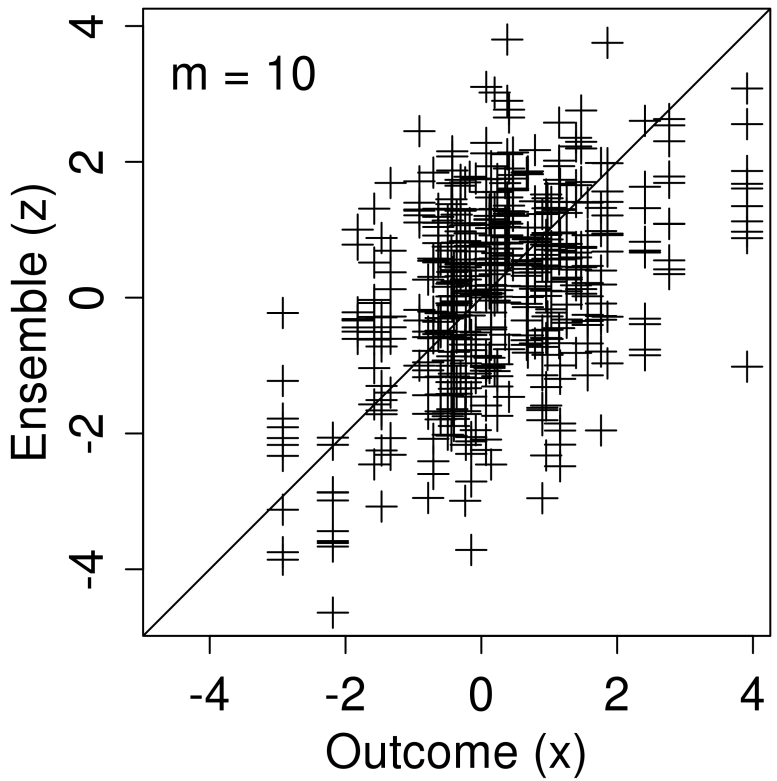
Let $F(t) = \Pr(x \leq t)$ be the probability forecast for $\mathbb{1}(x \leq t)$.

For ensemble $z_1, \ldots, z_m$ define $F(t) = \sum_{i=1}^{m} \mathbb{1}(z_i \leq t)/m$.

Example: Brier score,

$$\text{BS} = \{F(t) - \mathbb{1}(x \leq t)\}^2.$$

$\overline{\text{BS}} = 0.17$ (0.03) when $t = 0$.



Example: two proper scores

Let $F(t) = \Pr(x \leq t)$ be the probability forecast for $\mathbb{1}(x \leq t)$.

For ensemble $z_1, \ldots, z_m$ define $F(t) = \sum_{i=1}^{m} \mathbb{1}(z_i \leq t)/m$.

Example: Brier score,

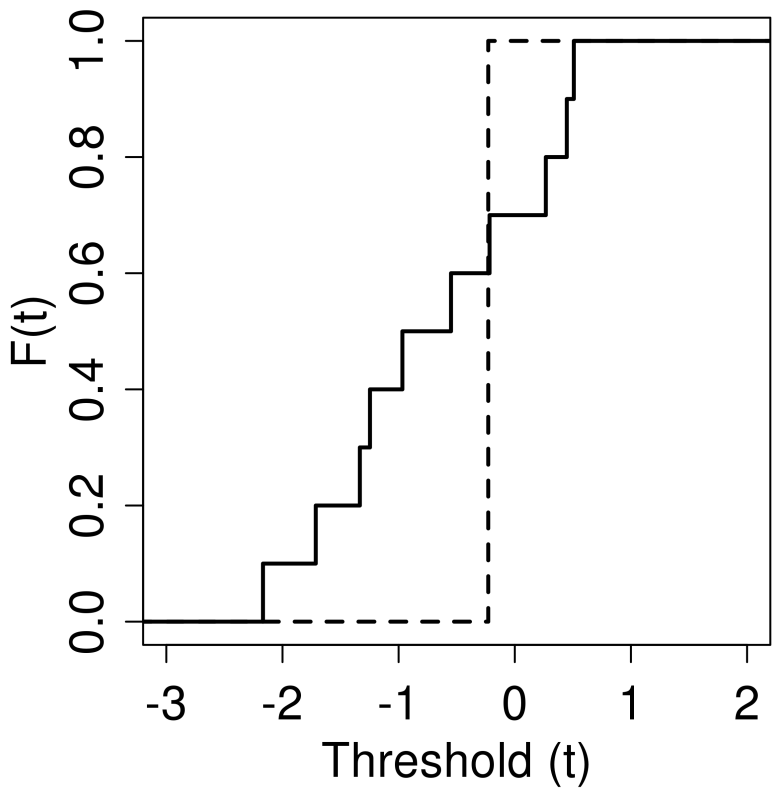
$$\text{BS} = \{F(t) - \mathbb{1}(x \leq t)\}^2.$$

$\overline{\text{BS}} = 0.17\ (0.03)$ when $t = 0$.

Example: Continuous Ranked Probability score,

$$\text{CRPS} = \int \{F(t) - \mathbb{1}(x \leq t)\}^2 \, \mathrm{d}t.$$

$\overline{\text{CRPS}} = 0.59\ (0.05)$.



Forecasts of some extremes are verified as usual

Forecasts of occurrences of extreme events (e.g. storms)

- Use proper scores for binary events, e.g. Brier score.

Forecasts of block maxima (e.g. annual maximum rainfall)

- Use proper scores for numerical outcomes, e.g. CRPS.

Forecasts of large values need care

Ask: how well do we forecast outcomes, x, that exceed u?

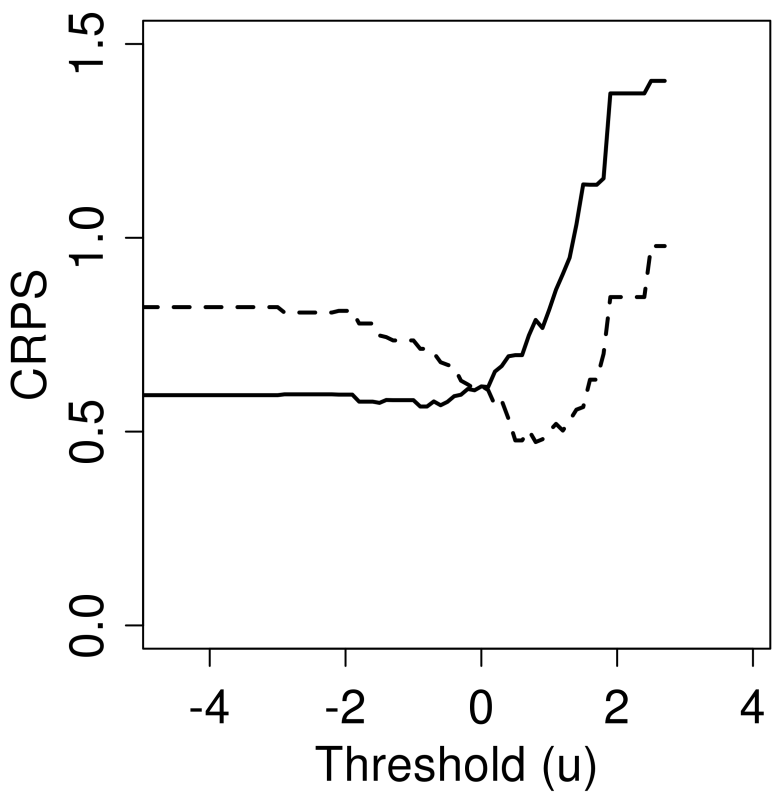
Suppose that we calculate a score using only cases with $x > u$.

This score is hedged by forecasts that assume $x > u$ always!

This phenomenon is called the forecaster's dilemma.

Example: $\overline{\text{CRPS}}$ calculated using only cases with $x > u$ for original forecasts (solid) and biased forecasts (dashed).

At high thresholds, the biased forecasts have better scores.



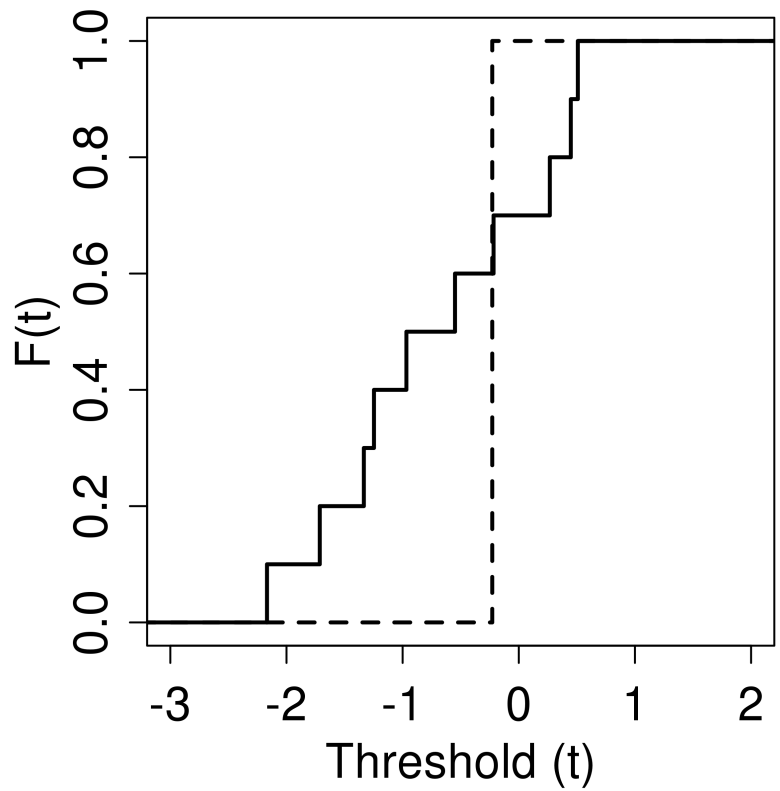
Use weighted scores for forecasts of large values

Should ask: how well do we forecast whether outcomes exceed u and, if they do, how well do we forecast the outcomes?

So we want good forecasts of $\Pr(x \leq u)$ and $\Pr(x \mid x > u)$.

Example: Threshold-weighted CRPS,

$$\int_u \{F(t) - \mathbb{1}(x \leq t)\}^2 \, \mathrm{d}t.$$



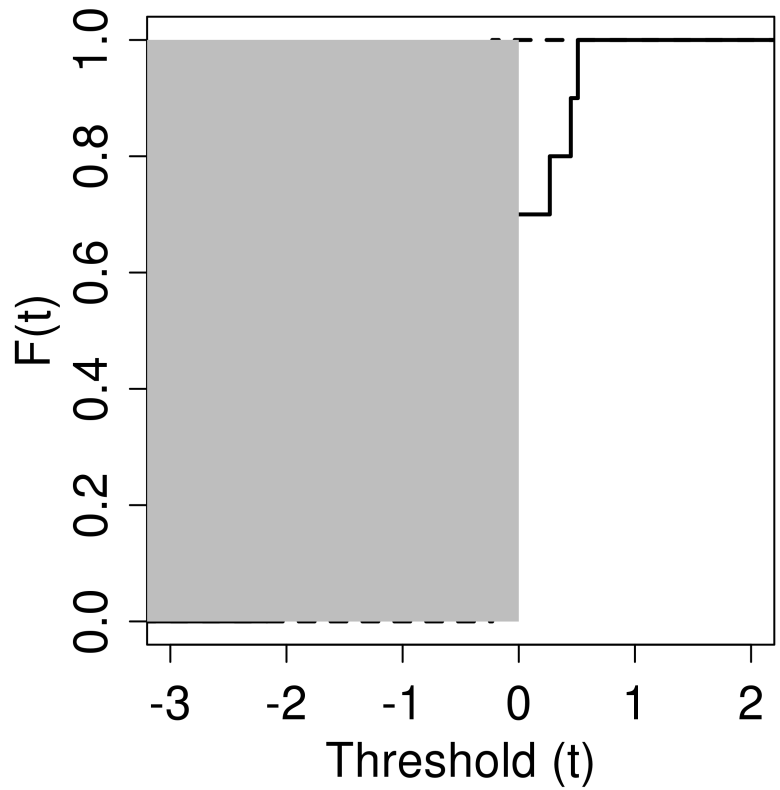
Use weighted scores for forecasts of large values

Should ask: how well do we forecast whether outcomes exceed u and, if they do, how well do we forecast the outcomes?

So we want good forecasts of $\Pr(x \leq u)$ and $\Pr(x \mid x > u)$.

Example: Threshold-weighted CRPS,

$$\int_u \{F(t) - \mathbb{1}(x \leq t)\}^2 \, \mathrm{d}t.$$



Use weighted scores for forecasts of large values

Should ask: how well do we forecast whether outcomes exceed u and, if they do, how well do we forecast the outcomes?

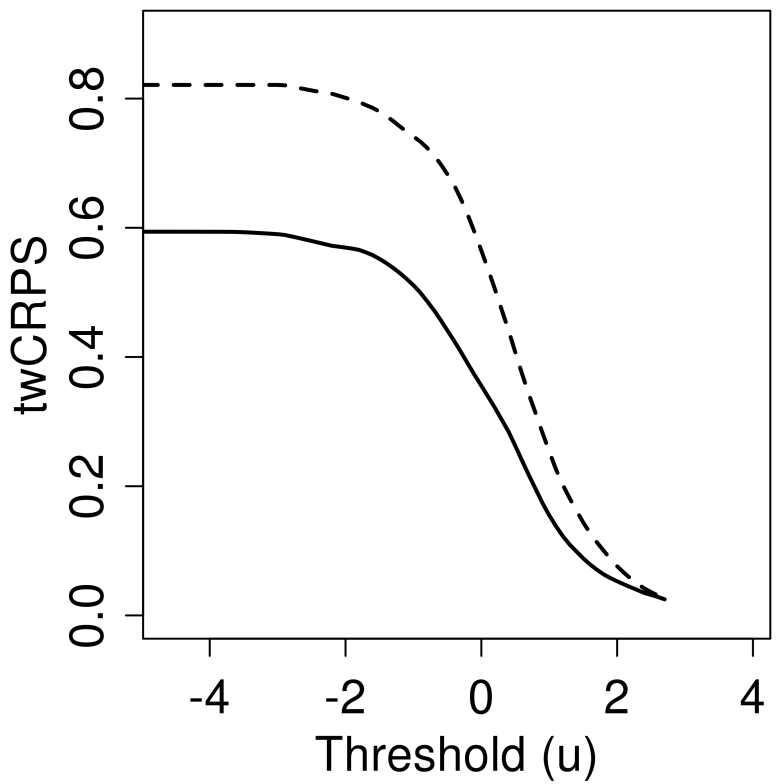
So we want good forecasts of $\Pr(x \leq u)$ and $\Pr(x \mid x > u)$.

Example: Threshold-weighted CRPS,

$$\int_u \{F(t) - \mathbb{1}(x \leq t)\}^2 \, \mathrm{d}t.$$

$\overline{\text{twCRPS}}$ for original (solid) and biased (dashed) forecasts.

Original forecasts have better scores at all thresholds.



Use fair scores to rank ensemble forecasts

We used proper scoring rules to verify ensemble probabilities.

This rewards an EPS if it produces good probabilities for its current ensemble size even if it would produce bad probabilities for other (e.g. infinite) ensemble sizes.

Example: Let $\bar{z}$ be the proportion of m ensemble members that forecast the event $\{x = 1\}$. If $m = 50$ and the event occurs 1% of the time then the long-run mean of the Brier score, $(\bar{z} - x)^2$, is optimized by ensembles that never forecast the event!

Use fair scores to rank ensemble forecasts

We used proper scoring rules to verify ensemble probabilities.

This rewards an EPS if it produces good probabilities for its current ensemble size even if it would produce bad probabilities for other (e.g. infinite) ensemble sizes.

Example: Let $\bar{z}$ be the proportion of m ensemble members that forecast the event $\{x = 1\}$. If $m = 50$ and the event occurs 1% of the time then the long-run mean of the Brier score, $(\bar{z} - x)^2$, is optimized by ensembles that never forecast the event!

Suppose that $x_1, x_2, \ldots$ have distribution p and that we sample ensembles $\mathbf{z}_1, \mathbf{z}_2, \ldots$ from one distribution, f, for all $x_1, x_2, \ldots$

Definition: A scoring rule, $s(\mathbf{z}, x)$, is **fair** if the long-run mean score is optimized by $f = p$.

Example: $(\bar{z} - x)^2$ is unfair; $(\bar{z} - x)^2 - \bar{z}(1 - \bar{z})/(m - 1)$ is fair.

Example: three fair scores

Let $F(t) = \Pr(x \leq t)$ be the probability forecast for $\mathbb{1}(x \leq t)$.

For ensemble $z_1, \ldots, z_m$ define $F(t) = m^{-1} \sum_{i=1}^{m} \mathbb{1}(z_i \leq t)$.

Example: Fair Brier score,

$$\text{BS} - \frac{F(t)\{1 - F(t)\}}{m - 1}.$$

Example: Fair CRPS,

$$\text{CRPS} - \int \frac{F(t)\{1 - F(t)\}}{m - 1} \, \mathrm{d}t.$$

Example: three fair scores

Let $F(t) = \Pr(x \leq t)$ be the probability forecast for $\mathbb{1}(x \leq t)$.

For ensemble $z_1, \ldots, z_m$ define $F(t) = m^{-1} \sum_{i=1}^{m} \mathbb{1}(z_i \leq t)$.

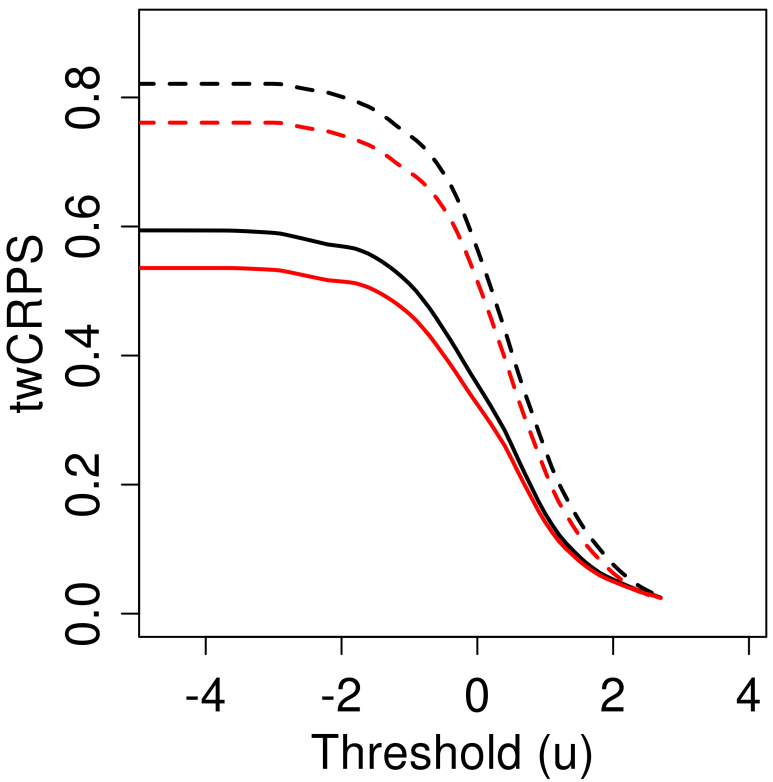
Example: Fair Brier score,

$$\text{BS} - \frac{F(t)\{1 - F(t)\}}{m - 1}.$$

Example: Fair CRPS,

$$\text{CRPS} - \int \frac{F(t)\{1 - F(t)\}}{m - 1} \, \mathrm{d}t.$$

Example: $\overline{\text{twCRPS}}$ (black) and fair $\overline{\text{twCRPS}}$ (red).



Adjust scores to the desired ensemble size

Fair scores are unbiased estimates of the scores that would be obtained if the ensemble size were infinite.

We also have unbiased estimates of the scores that would be obtained for any ensemble size, M.

Example: Adjusted Brier score,

$$\text{BS} - \frac{(1 - m/M)F(t)\{1 - F(t)\}}{m - 1}.$$

Example: Adjusted CRPS,

$$\text{CRPS} - \int \frac{(1 - m/M)F(t)\{1 - F(t)\}}{m - 1} \, \mathrm{d}t.$$

These can be used to predict the effects of changing ensemble size and to compare ensembles of different sizes.

Comparing forecasts of different events needs care

Proper scores tend to converge to the score for perfect forecasts as the predicted event becomes rarer.

Example: The long-run mean Brier score for climatological forecasts, p, is $p(1-p)$. This converges to 0 as $p \to 0$.

So we may need a lot of data to distinguish good forecasts.

This may not mean that forecasts of rarer events are better than forecasts of common events.

Use skill scores (the proportion of the maximum possible improvement over the reference forecast that is achieved) to compare forecasts of different events.

Example: Brier skill score $= 1 - \overline{\text{BS}}/\overline{\text{BS}}_{\text{ref}}$.

Summary

Use proper scores to rank probability forecasts.

Avoid calculating scores for only extreme outcomes.

Use weighted scores to focus on extreme outcomes.

Use (weighted) fair scores to rank ensemble forecasts.

Adjust scores to account for different ensemble sizes.

Avoid misinterpreting ‘better’ scores for rare events.

Scoring rules

Fricker, Ferro, Stephenson (2013) Three recommendations for evaluating climate predictions. Meteorological Applications, 20, 246–255

Gneiting, Raftery (2007) Strictly proper scoring rules, prediction, and estimation. Journal of the American Statistical Association, 102, 359–368

Weighted scoring rules

Gneiting, Ranjan (2011) Comparing density forecasts using threshold- and quantile-weighted scoring rules. Journal of Business & Economic Statistics, 29, 411–422

Lerch, Thorarinsdottir, Ravazzolo, Gneiting (2017) Forecaster's dilemma: extreme events and forecast evaluation. Statistical Science, 32, 106–127

Fair scoring rules

Ferro (2014) Fair scores for ensemble forecasts. QJRMS, 140, 1917–1923

Ferro, Richardson, Weigel (2008) On the effect of ensemble size on the discrete and continuous ranked probability scores. Meteorological Applications, 15, 19–24

Siegert, Ferro, Stephenson (2015) Evaluating ensemble forecasts by the ignorance score—correcting the finite-ensemble bias. arXiv:1410.8249v2

Degenerating scores

Ferro, Stephenson (2011) Extremal Dependence Indices: improved verification measures for deterministic forecasts of rare binary events. WAF, 26, 699–713

Related topics

Ferro (2007) A probability model for verifying deterministic forecasts of extreme events. Weather and Forecasting, 22, 1089–1100

Ferro (2017) Measuring forecast performance in the presence of observation error. Quarterly Journal of the Royal Meteorological Society, in press