What is a good ensemble forecast?

Chris Ferro

University of Exeter, UK

With thanks to Tom Fricker, Keith Mitchell, Stefan Siegert, David Stephenson, Robin Williams

6th International Verification Methods Workshop (17–19 March 2014, New Delhi)

What is a good ensemble forecast?

Chris Ferro

University of Exeter, UK

With thanks to Tom Fricker, Keith Mitchell, Stefan Siegert, David Stephenson, Robin Williams

6th International Verification Methods Workshop (17–19 March 2014, New Delhi)

What are ensemble forecasts?

An ensemble forecast is a set of possible values for the predicted quantity, often produced by running a numerical model with perturbed initial conditions.

A forecast's quality should depend on the meaning of the forecast that is claimed by the forecaster or adopted by the user, not the meaning preferred by the verifier.

For example, our assessment of a point forecast of 30°C for today's maximum temperature would depend on whether this is intended to be the value that will be exceeded with probability 50% or 10%.

What are ensembles *meant* to be?

Ensembles are intended to be simple random samples.

“the perturbed states are all equally likely”

— ECMWF website

“each member should be equally likely”

— Met Office website

“equally likely candidates to be the true state”

— Leith (1974)

How are ensembles *used*?

Turned into probability forecasts, e.g. proportion of members predicting a certain event

Turned into point forecasts, e.g. ensemble mean

What use is a random sample of possible values?

- inputs for an impact (e.g. hydrological) model
- easy for users to turn them into probability or point forecasts of their choice, e.g. sample statistics

How should we verify ensembles?

Should we verify...

probability forecasts derived from the ensemble,

point forecasts derived from the ensemble, or

the ensemble forecast as a random sample?

The first two approaches are more common, but ensembles that do well as random samples need not produce the best probability or point forecasts.

Results depend on what you verify

The ideal probability forecast should behave as if the verification is sampled from the forecast distribution (Diebold et al. 1998).

The ideal ensemble forecast should behave as if it and the verification are sampled from the same distribution (Anderson 1997).

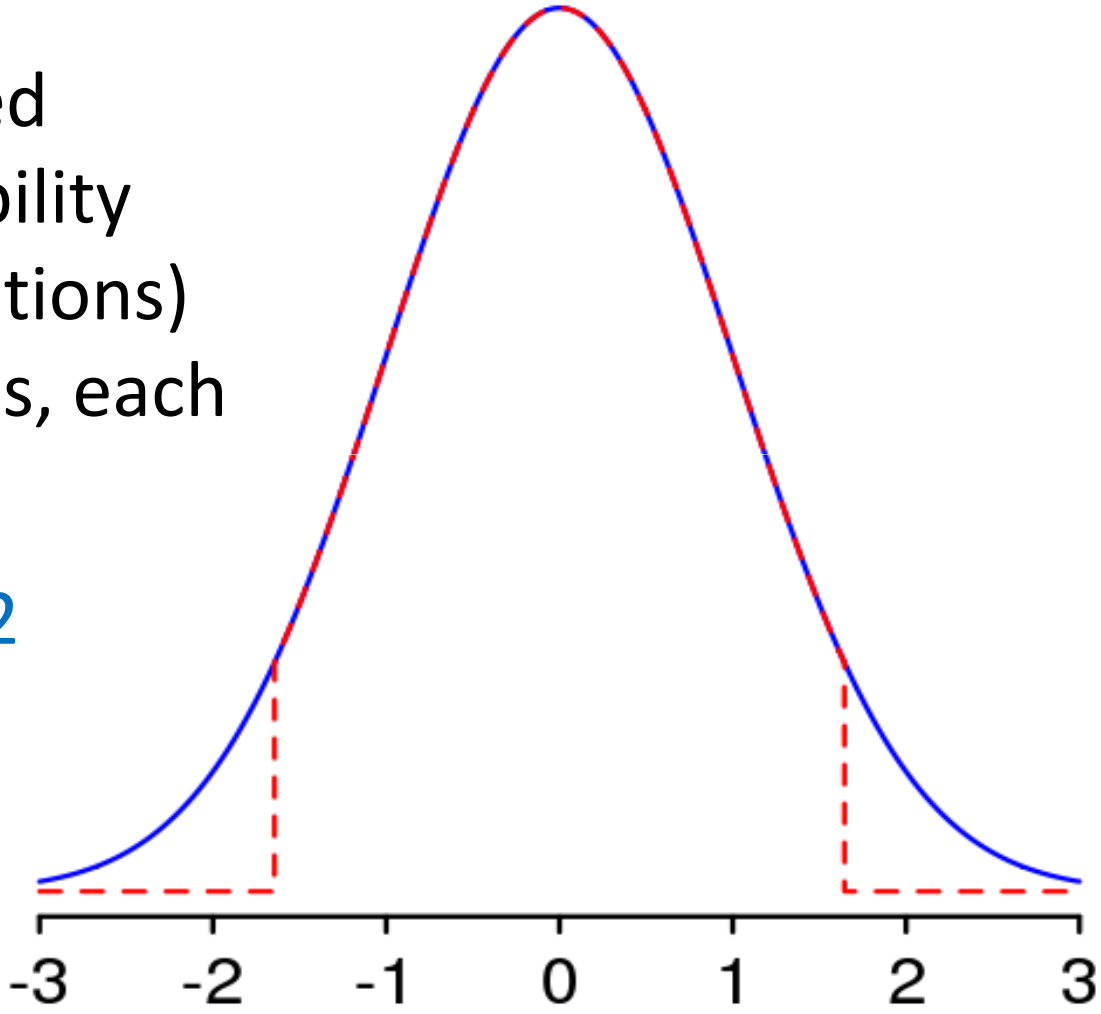
An ideal ensemble will not appear optimal if we verify a derived probability forecast because the latter will only *estimate* the ideal probability forecast.

An illustrative example

Expected Continuous Ranked Probability Score for probability forecasts (empirical distributions) derived from two ensembles, each with ten members.

Ideal ensemble: CRPS = 0.62

Underdispersed ensemble: CRPS = 0.61 (better!)



How should we verify ensembles?

Verify an ensemble according to its intended use.

If we want its empirical distribution to be used as a probability forecast then we should verify the empirical distribution with a proper score.

If we want its ensemble mean to be used as a mean point forecast then we should verify the ensemble mean with the squared error.

If we want the ensemble to be used as a random sample then what scores should we use?

Fair scores for ensembles

Fair scores favour ensembles that behave as if they and the verification are sampled from the same distribution.

A score, s(x,y), for an ensemble forecast, x, sampled from p, and a verification, y, is *fair* if (for all p) the expected score, $E_{x,y}\{s(x,y)\}$, is optimized when y ~ p.

In other words, the expected value of the score over all possible ensembles is a proper score.

Fricker et al. (2013)

Characterization: binary case

Let y = 1 if an event occurs, and let y = 0 otherwise.

Let $s_{i,y}$ be the (finite) score when i of m ensemble members forecast the event and the verification is y.

The (negatively oriented) score is fair if

$$(m - i)(s_{i+1,0} - s_{i,0}) = i(s_{i-1,1} - s_{i,1})$$

for i = 0, 1, ..., m and $s_{i+1,0} \geq s_{i,0}$ for i = 0, 1, ..., m – 1.

Ferro (2013)

Examples of fair scores

Verification y and m ensemble members x_1, ..., x_m.

The fair Brier score is

$$s(x,y) = (i/m - y)^2 - i(m - i)/\{m^2(m - 1)\}$$

where i members predict the event {y = 1}.

The fair CRPS is

$$s(x,y) = \int \{i(t)/m - H(t - y)\}^2 dt - \sum_{i \neq j} |x_i - x_j| / \{2m^2(m - 1)\}$$

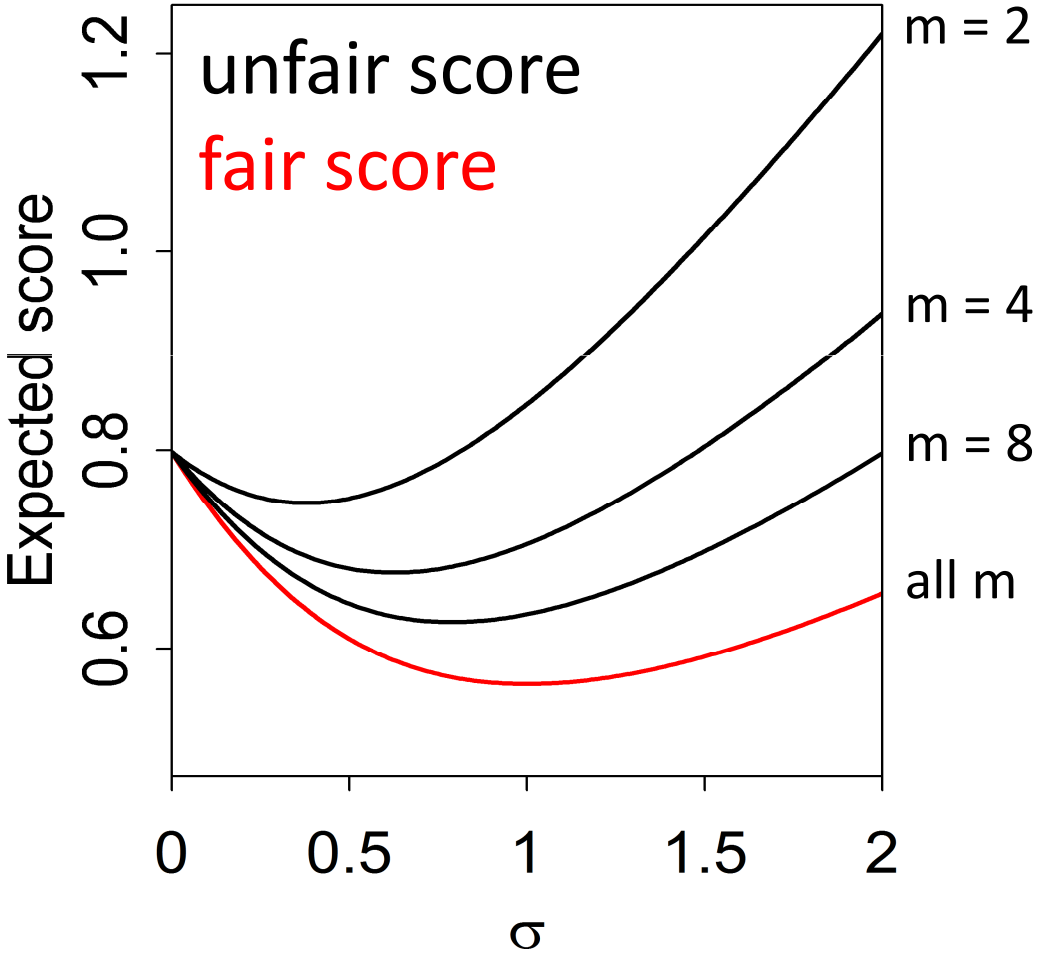
where i(t) members predict the event {y ≤ t} and H is the Heaviside function.

Example: CRPS

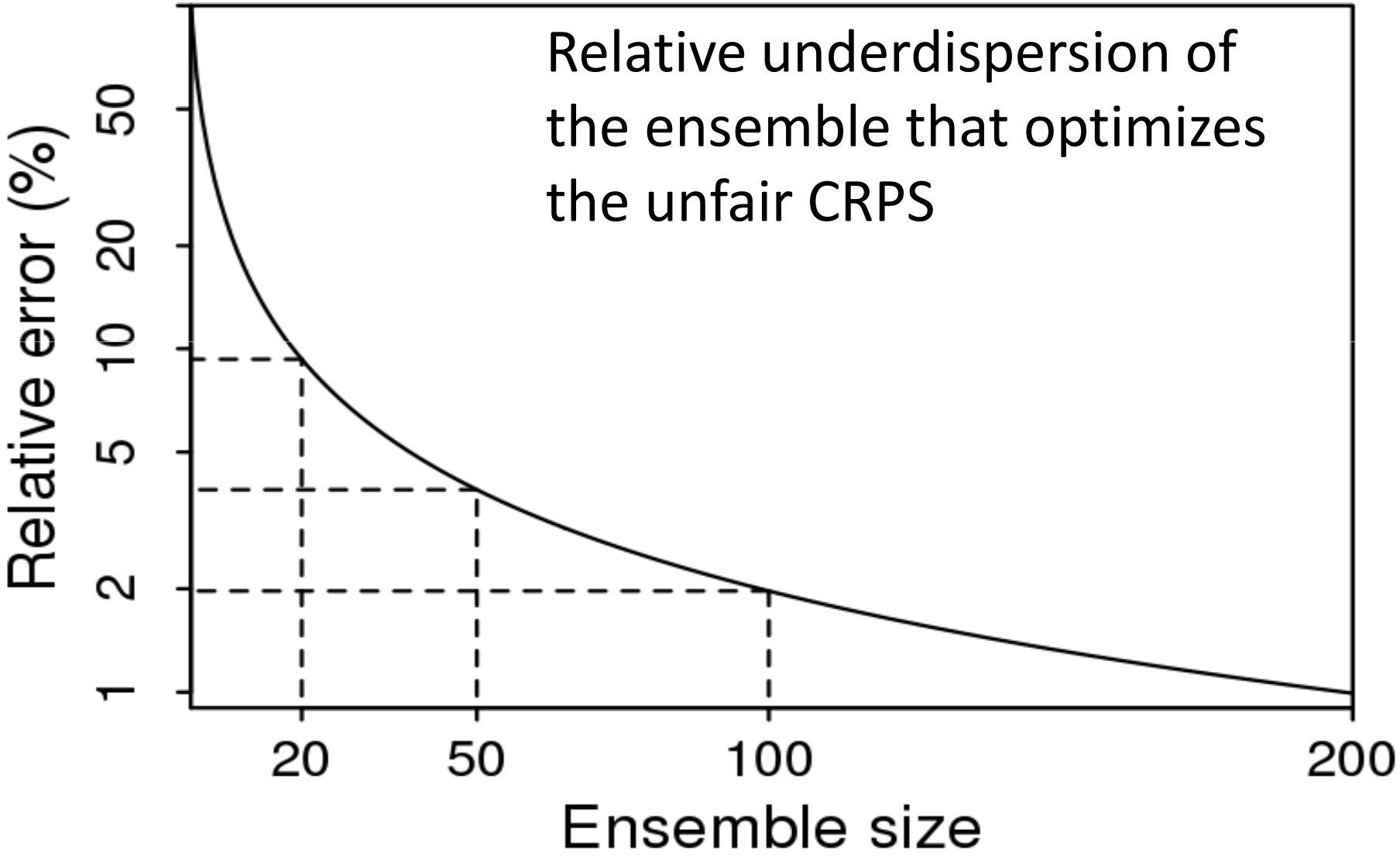
Verifications $y \sim N(0,1)$ and m ensemble members $x_i \sim N(0,\sigma^2)$ for $i = 1, \ldots, m$.

Expected values of the fair and unfair CRPS against σ.

The fair CRPS is always optimized when ensemble is well dispersed ($\sigma = 1$).



Example: CRPS



If the verification is an ensemble?

A score, s(p,y), for a probability forecast, p, and an n-member ensemble verification, y, is *n-proper* if (for all p) the expected score, $E_y\{s(p,y)\}$, is optimized when $y_1, ..., y_n \sim p$ (Thorarinsdottir et al. 2013).

A score, s(x,y), for an m-member ensemble forecast, x, sampled from p, and an n-member ensemble verification, y, is *m/n-fair* if (for all p) the expected score, $E_{x,y}\{s(x,y)\}$, is optimized when $y_1, ..., y_n \sim p$.

If the verification is an ensemble?

If $s(x,y_i)$ is (m-)fair then $n^{-1}\sum_{i=1}^{n} s(x, y_i)$ is m/n-fair.

The m/n-fair Brier score is

$$s(x,y) = (i/m - j/n)^2 - i(m - i)/\{m^2(m - 1)\}$$

where i members and j verifications predict {y = 1}.

The m/n-fair CRPS is

$$s(x, y) = \int \{i(t)/m - j(t)/n\}^2 dt - \sum_{i \neq j} |x_i - x_j| / \{2m^2(m - 1)\}$$

where i(t) members and j(t) verifications predict {y ≤ t}.

Summary

Verify an ensemble according to its intended use.

Verifying derived probability or point forecasts will not favour ensembles that behave as if they and the verifications are sampled from the same distribution.

Fair scores should be used to verify ensemble forecasts that are to be interpreted as random samples.

Fair scores are analogues of proper scores, and rank histograms are analogues of PIT histograms.

Current work: analogues of reliability diagrams etc.

References

Anderson JL (1997) The impact of dynamical constraints on the selection of initial conditions for ensemble predictions: low-order perfect model results. *Monthly Weather Review*, 125, 2969-2983

Diebold FX, Gunther TA, Tay AS (1998) Evaluating density forecasts with applications to financial risk management. *International Economic Review*, 39, 863-882

Ferro CAT (2013) Fair scores for ensemble forecasts. *Quarterly Journal of the Royal Meteorological Society*, in press

Fricker TE, Ferro CAT, Stephenson DB (2013) Three recommendations for evaluating climate predictions. *Meteorological Applications*, 20, 246-255

Leith CE (1974) Theoretical skill of Monte Carlo forecasts. *Monthly Weather Review*, 102, 409-418

Thorarinsdottir TL, Gneiting T, Gissibl N (2013) Using proper divergence functions to evaluate climate models. *SIAM/ASA J. on Uncertainty Quantification*, 1, 522-534

Weigel AP (2012) Ensemble forecasts. In *Forecast Verification: A Practitioner's Guide in Atmospheric Science*. IT Jolliffe, DB Stephenson (eds) Wiley, pp. 141-166