Ensemble verification: Old scores, new perspectives

Sabrina Wahl, Petra Friederichs, Jan Keller

Hans-Ertel-Zentrum für Wetterforschung

Meteorologisches Institut Universität Bonn

universitätbonn

WMO Verification Workshop
Berlin, May 2017

ensemble forecast

equally probable simulations of numerical model

- ➡ calibration: rank (pit) histogram, beta score
- ➡ discrimination: generalized discrimination score
- ➡ sharpness: prediction interval

translation, interpretation, post-processing

probabilistic forecast

pdf, cdf, mean, sd, quantiles, probabilities,…

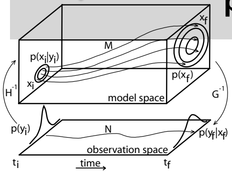
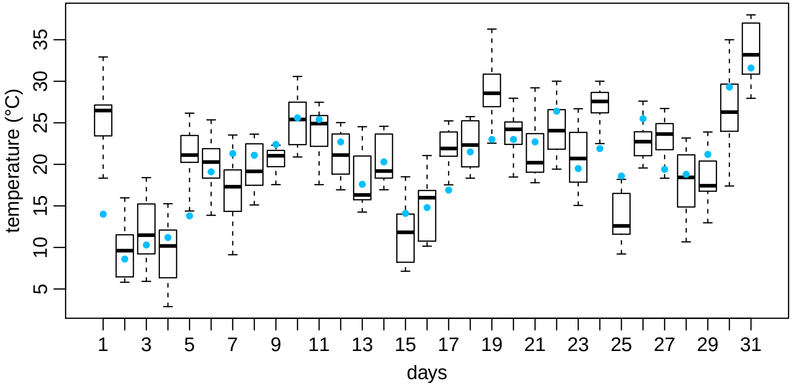


Fig. 1 from Stephenson et al. (2005)

- ➡ proper scoring rules:
  CRPS, Brier score, quantile score, logarithmic score, MSE, MAE, …



- ensemble forecast in terms of empirical distribution
- boxplot represents forecast distribution in terms of quantiles
- evaluation of ensemble members as quantiles

Verification-framework for quantiles

- score for quantile forecasts q_τ when y is the event that materializes, with $\tau \in (0,1)$ the probability level

$$S_Q(q_\tau, y) = \rho_\tau(y - q_\tau) = \begin{cases} \mid y - q_\tau \mid \tau & \text{if } y \geq q_\tau \\ \mid y - q_\tau \mid (1-\tau) & \text{if } y < q_\tau \end{cases}$$

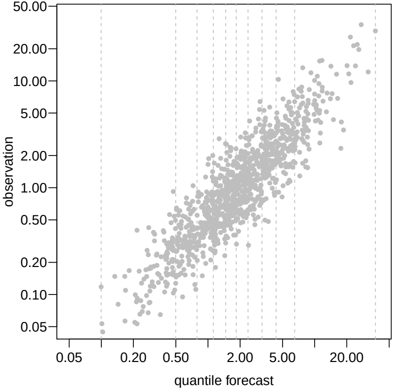
- empirical quantile score from a set of N forecast-observation pairs

$$QS(\tau) = \frac{1}{N} \sum_{i=1}^{N} \rho_\tau(y_i - q_{\tau,i})$$

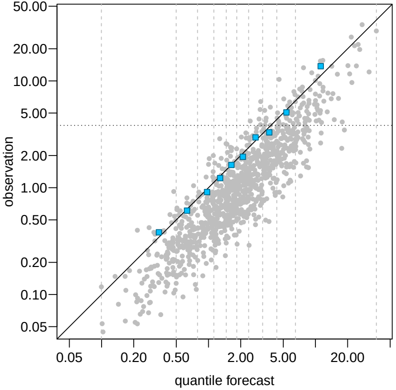
- decomposition of the quantile score *(Bentzien and Friederichs, 2014)*

$$QS(\tau) = \frac{1}{N} \sum_{i=1}^{N} \rho_\tau(y_i - q_{\tau,i}) = UNC(\tau) - RES(\tau) + REL(\tau)$$

- **Calibration**: quantile reliability diagram
- forecast intervals $\mathcal{I}_k$
- $y_\tau^{(k)}$: conditional observed quantile in $\mathcal{I}_k$
- discrete values $y_\tau^{(k)}, q_\tau^{(k)}$ with $k = 1, ..., K \leq N$



- **Calibration**: quantile reliability diagram
- forecast intervals $\mathcal{I}_k$
- $y_\tau^{(k)}$: conditional observed quantile in $\mathcal{I}_k$
- discrete values $y_\tau^{(k)}, q_\tau^{(k)}$ with $k = 1, ..., K \leq N$



- **Reliability**, perfect if $y_\tau^{(k)} = q_\tau^{(k)}$

$$REL = \frac{1}{N} \sum_{k=1}^{K} \sum_{n \in \mathcal{I}_k} \left[\rho_\tau \left(y_n - q_\tau^{(k)} \right) - \rho_\tau \left(y_n - \bar{y}_\tau^{(k)} \right) \right]$$

- **Resolution**, good if $y_\tau^{(k)} \neq \bar{y}_\tau$

$$RES = \frac{1}{N} \sum_{k=1}^{K} \sum_{n \in \mathcal{I}_k} \left[\rho_\tau \left(y_n - \bar{y}_\tau \right) - \rho_\tau \left(y_n - \bar{y}_\tau^{(k)} \right) \right]$$

- **Uncertainty**, from sample climatology $\bar{y}_\tau$

$$UNC = \frac{1}{N} \sum_{n=1}^{N} \rho_\tau (y_n - \bar{y}_\tau)$$

- Score for multiple quantiles $q_{\tau_1}, ..., q_{\tau_k}$ with $\tau_1, ..., \tau_k \in (0, 1)$

$$S_Q(q_{\tau_1}, ..., q_{\tau_k}, y) = \sum_{i=1}^{k} \rho_{\tau_i}(y - q_{\tau_i})$$

- interpret ensemble members $e^{(1)} \leq e^{(2)} \leq ... \leq e^{(M)}$ as quantiles to the probability levels $\tau_1, ..., \tau_M \in (0, 1)$

$$QS_{ENS} = \sum_{j=1}^{M} QS(\tau_j) = \sum_{j=1}^{M} \left[\frac{1}{N} \sum_{i=1}^{N} \rho_{\tau_j} \left(y_i - e_i^{(j)} \right) \right]$$

- quantile score decomposition for ensemble

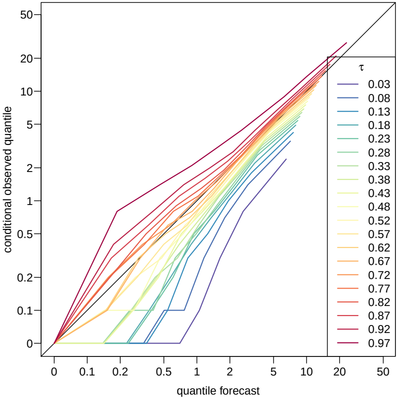
$$QS_{ENS} = \sum_{j=1}^{M} UNC(\tau_j) - \sum_{j=1}^{M} RES(\tau_j) + \sum_{j=1}^{M} REL(\tau_j)$$

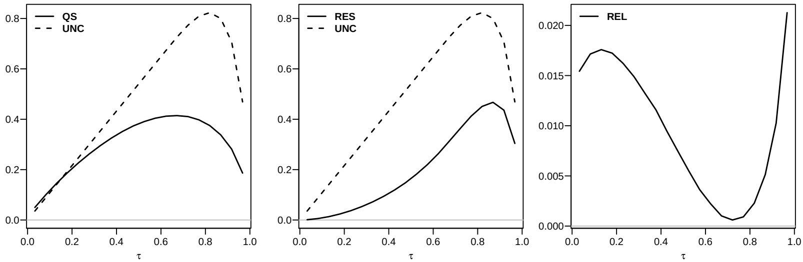
$$QS_{ENS} = \sum_{j=1}^{M} UNC(\tau_j) - \sum_{j=1}^{M} RES(\tau_j) + \sum_{j=1}^{M} REL(\tau_j)$$

- quantile reliability curves for each τ_j
- graphical exploration of $UNC(\boldsymbol{\tau}), RES(\boldsymbol{\tau}), REL(\boldsymbol{\tau})$ for $\boldsymbol{\tau} = (\tau_1, ..., \tau_M)$

Example:
COSMO-DE-EPS 12-hourly precipitation forecasts for 365 days in 2011.
Number of observations $N = 384\,679$ (from 1079 observing sites).
Number of ensemble members $M = 20$.

- quantile reliability curves should be close to diagonal
- ”spread” around the diagonal indicates insufficient ensemble spread
- underestimation of higher quantiles
- overestimation of lower quantiles





- graphical exploration of $UNC(\tau)$, $RES(\tau)$, $REL(\tau)$
- optimal score:

$$QS = 0$$
$$REL = 0$$
$$RES = UNC$$

- quantile score decomposition

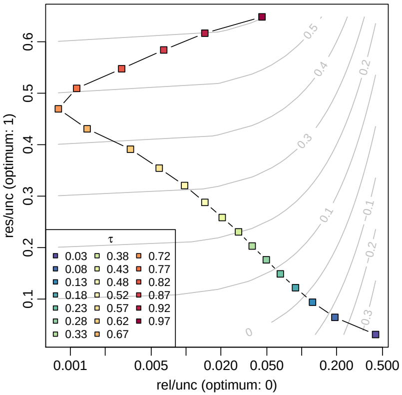
$$QS = UNC - RES + REL \tag{1}$$

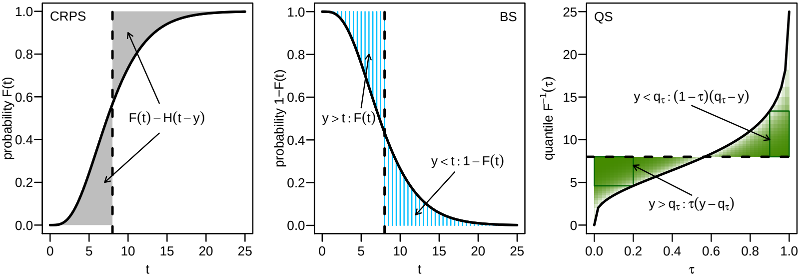
- uncertainty is independent of forecasts, divide eq. (1) by UNC

$$QSS = 1 - \frac{QS}{UNC} = \frac{RES}{UNC} - \frac{REL}{UNC} \tag{2}$$

- optimal values
  - $QSS = 1$ maximum improvement over climatology
  - $RES/UNC = 1$ maximum achievable resolution
  - $REL/UNC = 0$ perfect calibration

- plot scaled resolution vs. scaled reliability
- contours show lines of constant quantile skill score
- combine three forecast attributes in one diagram
- compare different quantiles and/or forecast models





$$S_{CRP} = \int_{\mathcal{R}} S_B(1 - F(u), y)\, du = 2 \int_0^1 S_Q(F^{-1}(\tau), y)\, d\tau$$

see e.g. *Gneiting and Raftery (2007)*

- let $e^{(1)} \leq e^{(2)} \leq ... \leq e^{(M)}$ be an ensemble forecast for Y
- cumulative distribution function from ensemble

$$F_e(x) = \sum_{i=1}^{M} w_i \, H(x - e^{(i)})$$

- weights $w_i > 0$ and $\sum_{i=1}^{M} w_i = 1$
- F_e features exactly M jumps at the points $x = e^{(i)}$ with jump height w_i

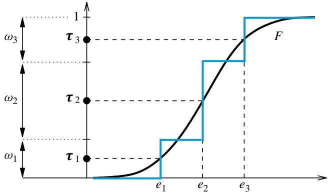


Fig. 1 from *Broecker (2012)*

- score for distribution F_e

$$S_{CRP}(F_e, y) = \int [F_e(x) - H(x-y)]\, dx$$

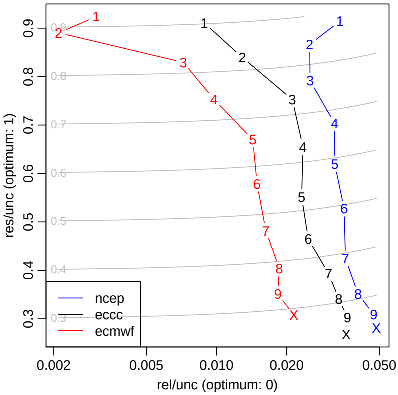
- is equivalent to sum of weighted quantile scores *(Broecker, 2012)*

$$S_{CRP}(F_e, y) = 2\sum_{i=1}^{M} w_i\, \rho_{\tau_i}(y - e^{(i)})$$

- with decomposition

$$S_{CRP}(F_e, y) = 2\sum_{i=1}^{M} w_i\, UNC(\tau_i) - 2\sum_{i=1}^{M} w_i\, RES(\tau_i) + 2\sum_{i=1}^{M} w_i\, REL(\tau_i)$$

- contours show lines of constant CRPS skill score
- scaled resolution and reliability: sum over all τ
- compare different forecast models and/or lead times



Example:
Global EPS daily 12 UTC 500 hPa geopotenial forecasts for 30 days in 2012 (JJA).
Number of gridboxes: 720 $\times$ 361 (observations: ERA Interim).
Number of ensemble members: 20 to 50.

Summary

$$S_{CRP}(F_e, y) = 2\sum_{i=1}^{M} w_i\, UNC(\tau_i) - 2\sum_{i=1}^{M} w_i\, RES(\tau_i) + 2\sum_{i=1}^{M} w_i\, REL(\tau_i)$$

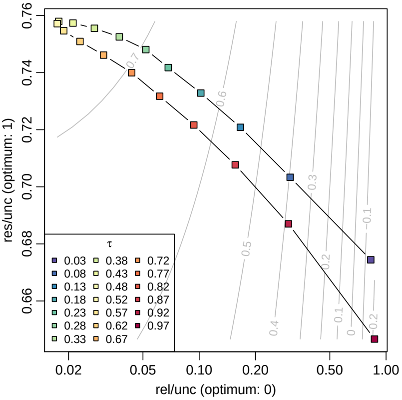
- **Ensemble verification** using quantiles can have different levels of complexity
- Representation of CRPS as weighted sum over quantile scores
  1. CRPS single value (compare different models, lead times, ...)
  2. CRPS attributes: skill, resolution and reliability as function of τ
  3. quantile reliability curves
- Application to empirical distribution as well as to parametric distribution derived from statistical postprocessing

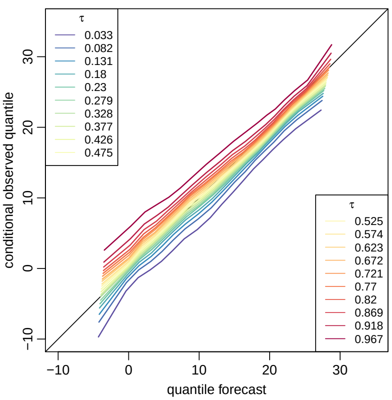
1. Bentzien and Friederichs, “Decomposition and graphical portrayal of the quantile score,” Quarterly Journal of the Royal Meteorological Society, vol. 140, pp. 1924–1934, 2014.
2. Broecker, ”Evaluating raw ensembles with the continuous ranked probability score”, Quarterly Journal of the Royal Meteorological Society, vol. 138, pp. 1611–1617, 2012.
3. Gneiting and Raftery, “Strictly proper scoring rules, prediction, and estimation”, Journal of the American Statistical Association, vol. 102, pp. 359–378, 2007.
4. Hyndman and Fan, ”Sample quantiles in statistical packages”, The American Statistician, vol. 50, pp. 361-365, 1996.
5. Stephenson et al., ”Forecast assimilation: a unified framework for the combination of multi-model weather and climate predictions”, Tellus, vol. 57 pp. 253-264, 2005.

Hyndman and Fan (1996): Sample quantiles in statistical packages

- *Definition 4:* $\tau_j = \frac{j}{M}$
- *Definition 5:* $\tau_j = \frac{j-0.5}{M}$
- *Definition 6:* $\tau_j = \frac{j}{M+1}$
- *Definition 7:* $\tau_j = \frac{j-1}{M-1}$
- *Definition 8:* $\tau_j = \frac{j-1/3}{M+1/3}$
- *Definition 9:* $\tau_j = \frac{j-3/8}{M+1/4}$

for $j = 1, ..., M$ (number of ensemble members)





Example:
COSMO-DE-EPS daily 12 UTC temperature forecasts for 365 days in 2011.
Number of observations $N = 174\,603$ (from 481 observing sites).
Number of ensemble members $M = 20$.