Ensemble Verification I

Renate Hagedorn

European Centre for Medium-Range Weather Forecasts

Objective of diagnostic/verification tools

Assessing the *goodness* of a forecast system involves determining skill and value of forecasts

A forecast has skill if it predicts the observed conditions well according to some objective or subjective criteria.

A forecast has value if it helps the user to make better decisions than without knowledge of the forecast.

- Forecasts with poor skill can be valuable (e.g. location mismatch)
- Forecasts with high skill can be of little value (e.g. blue sky desert)

Assessing the quality of a forecast system

- Characteristics of a forecast system:
  - **Consistency***: Do the observations statistically belong to the distributions of the forecast ensembles? (consistent degree of ensemble dispersion)
  - **Reliability**: Can I trust the probabilities to mean what they say?
  - **Sharpness**: How much do the forecasts differ from the climatological mean probabilities of the event?
  - **Resolution**: How much do the forecasts differ from the climatological mean probabilities of the event, and the systems gets it right?
  - **Skill**: Are the forecasts better than my reference system (chance, climatology, persistence,...)?

* Note that terms like consistency, reliability etc. are not always well defined in verification theory and can be used with different meanings in other contexts

Assessing the quality of a forecast system

- Characteristics of a forecast system:

Rank Histogram

- **Consistency**: Do the observations statistically belong to the distributions of the forecast ensembles? (consistent degree of ensemble dispersion)

Reliability Diagram

- **Reliability**: Can I trust the probabilities to mean what they say?
- **Sharpness**: How much do the forecasts differ from the climatological mean probabilities of the event?
- **Resolution**: How much do the forecasts differ from the climatological mean probabilities of the even, and the systems gets it right?

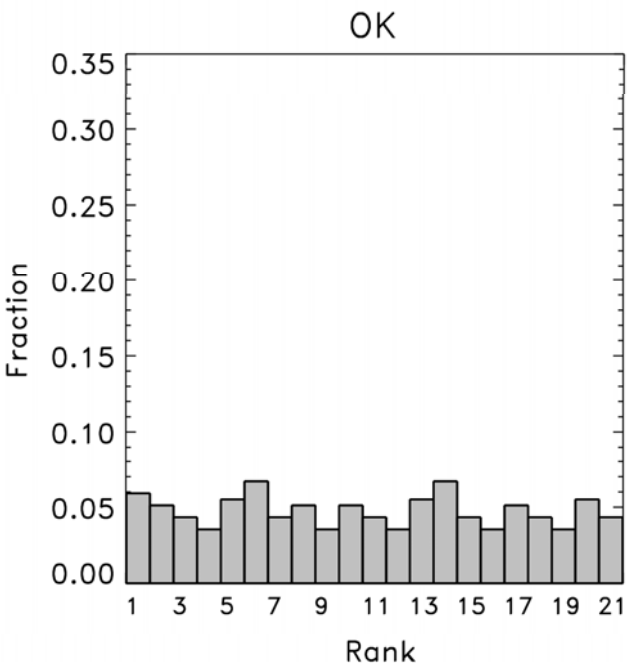
Brier Skill Score

- **Skill**: Are the forecasts better than my reference system (chance, climatology, persistence,...)?

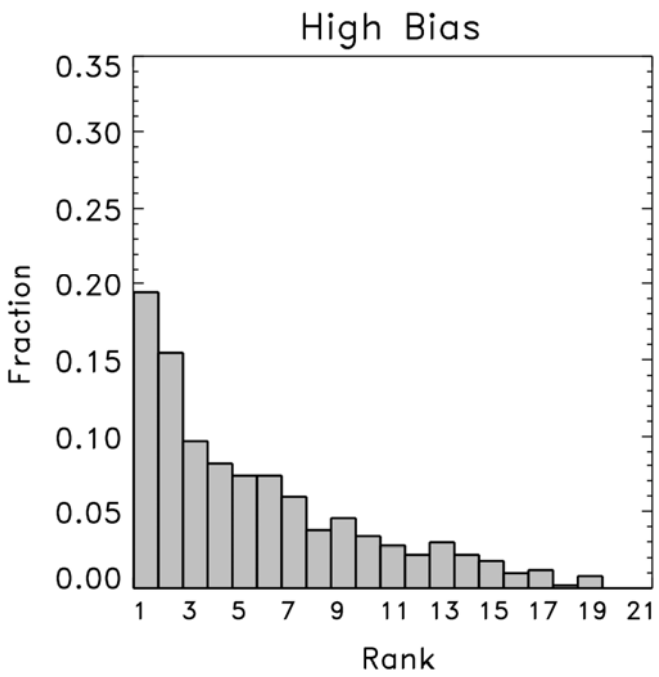
Rank Histogram

- Rank Histograms asses whether the ensemble spread is consistent with the assumption that the observations are statistically just another member of the forecast distribution
  - Check whether observations are equally distributed amongst predicted ensemble
  - Sort ensemble members in increasing order and determine where the observation lies with respect to the ensemble members

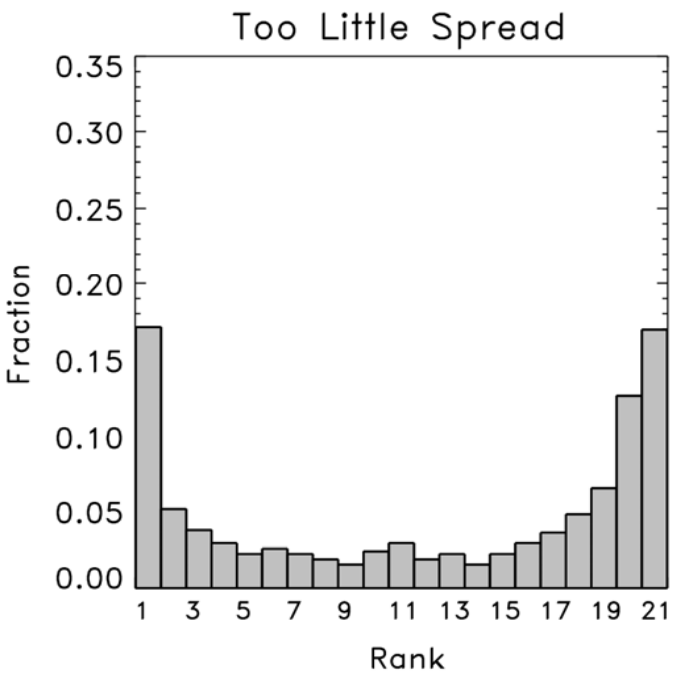
Rank Histograms



OBS is indistinguishable from any other ensemble member



OBS is too often below the ensemble members (biased forecast)

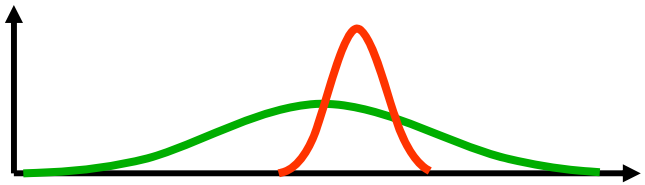


OBS is too often outside the ensemble spread

A uniform rank histogram is a necessary but not sufficient criterion for determining that the ensemble is reliable (see also: T. Hamill, 2001, MWR)

Reliability

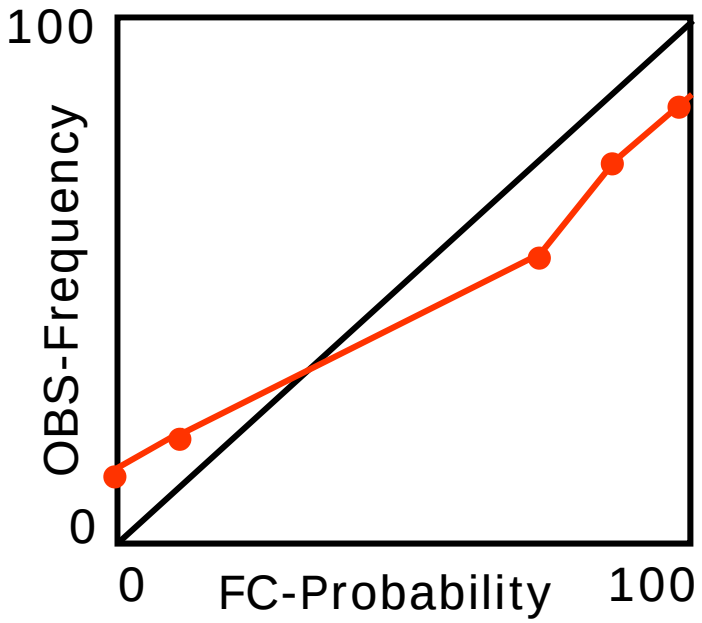
- A forecast system is reliable if:
  - statistically the predicted probabilities agree with the observed frequencies, i.e.
  - taking all cases in which the event is predicted to occur with a probability of x%, that event should occur exactly in x% of these cases; not more and not less.
- A reliability diagram displays whether a forecast system is reliable (unbiased) or produces over-confident / under-confident probability forecasts
- A reliability diagram also gives information on the resolution (and sharpness) of a forecast system



Forecast PDF
Climatological PDF

Reliability Diagram

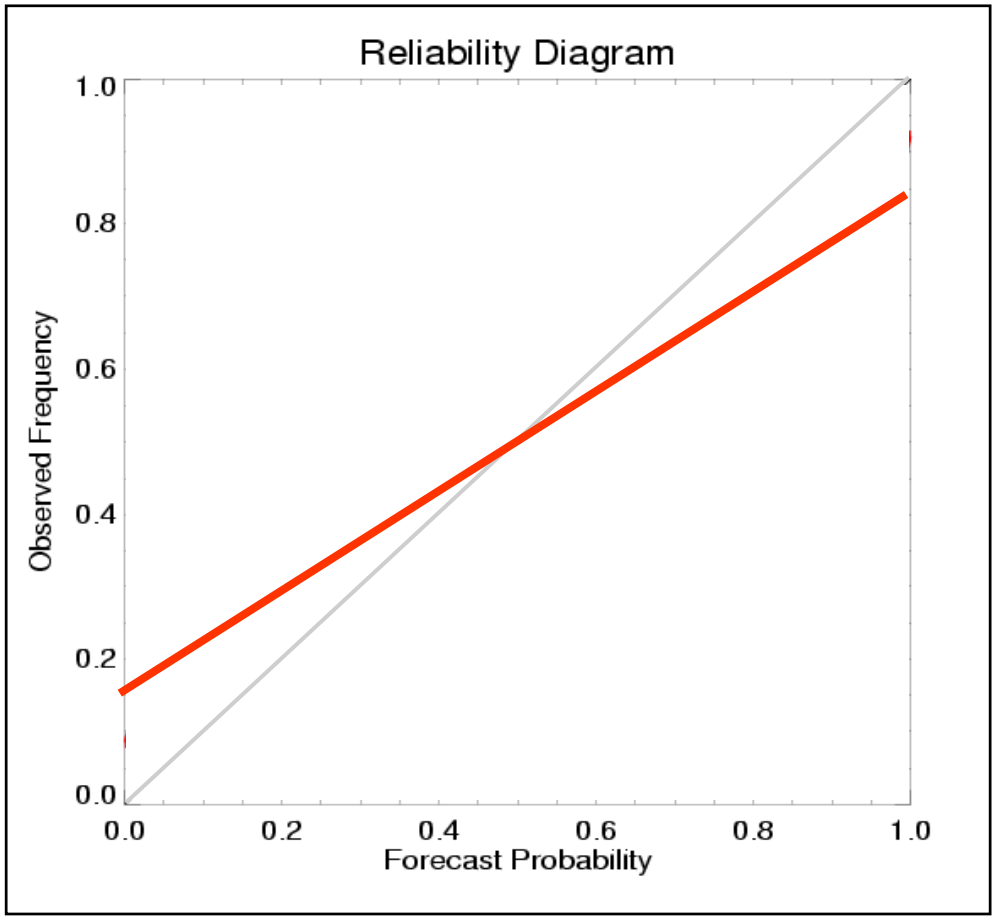
Take a sample of probabilistic forecasts:
e.g. 30 days x 2200 GP = 66000 forecasts
How often was event (T > 25) forecasted with X probability?



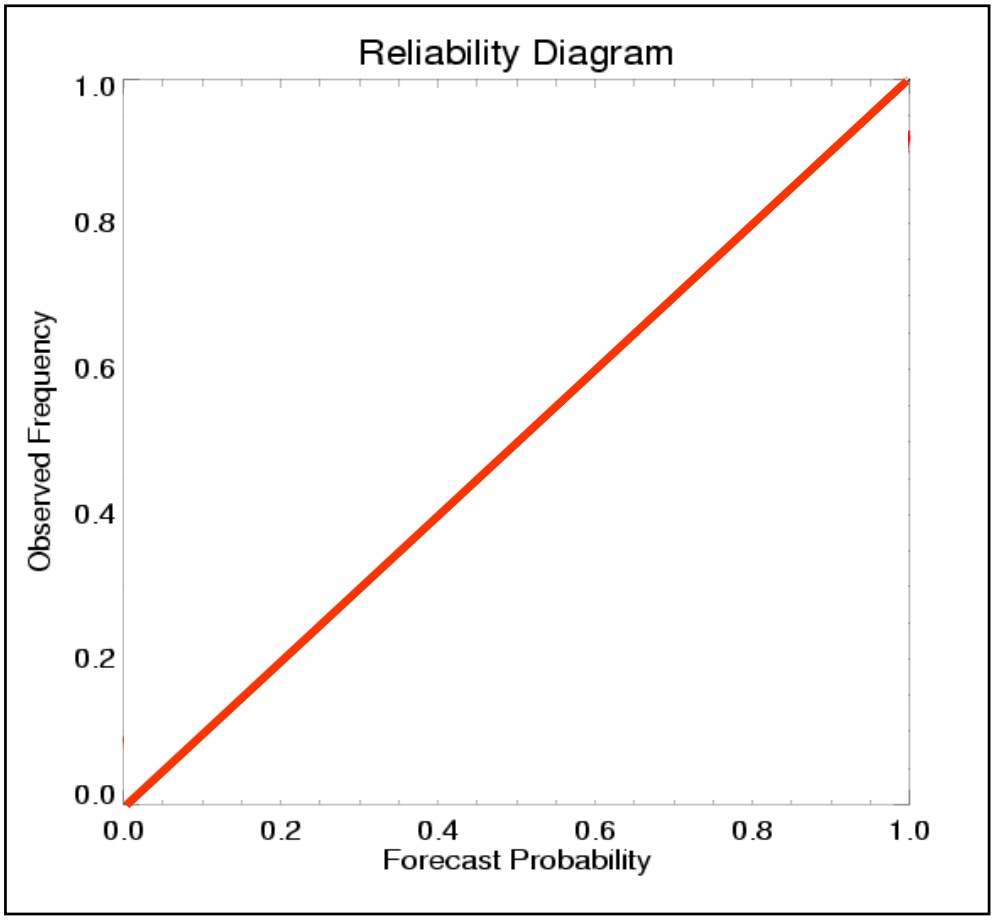
FC Prob.	# FC	"perfect FC" OBS-Freq.	"real" OBS-Freq.
100%	8000	8000 (100%)	7200 (90%)
90%	5000	4500 (90%)	4000 (80%)
80%	4500	3600 (80%)	3000 (66%)
….	….	….	….
….	….	….	….
….	….	….	….
10%	5500	550 (10%)	800 (15%)
0%	7000	0 (0%)	700 (10%)

Reliability Diagram

over-confident model

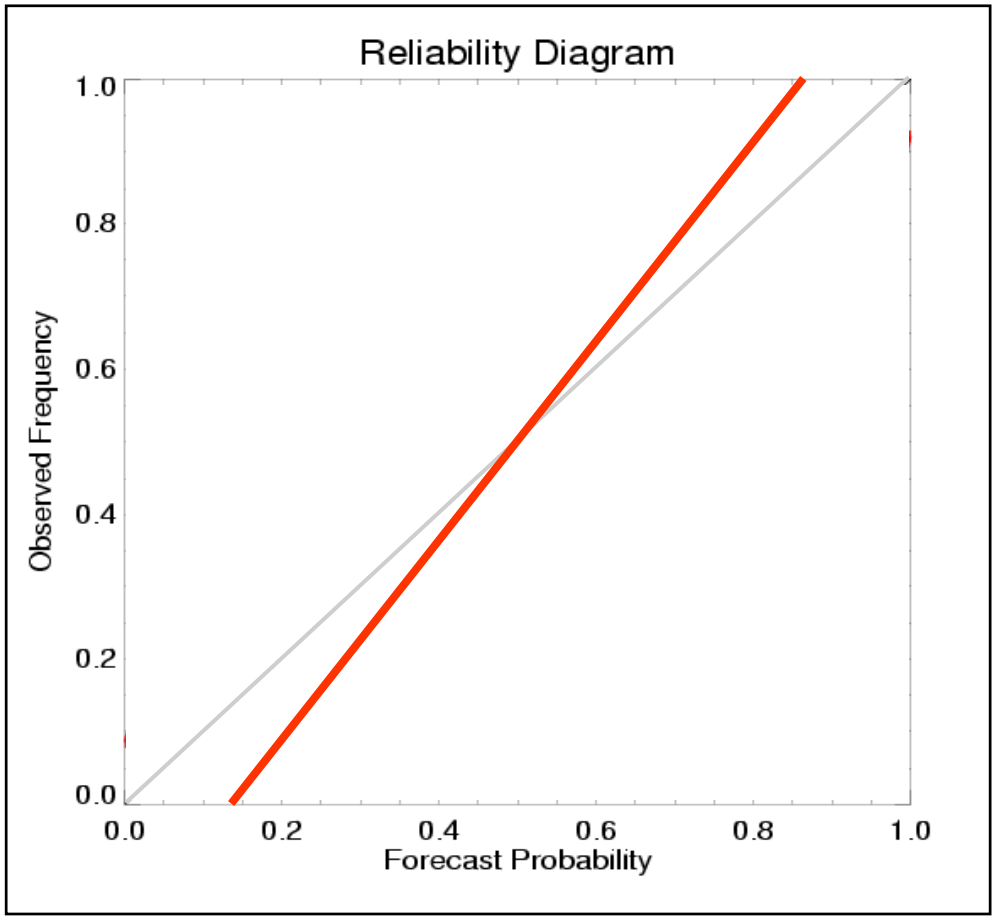


perfect model

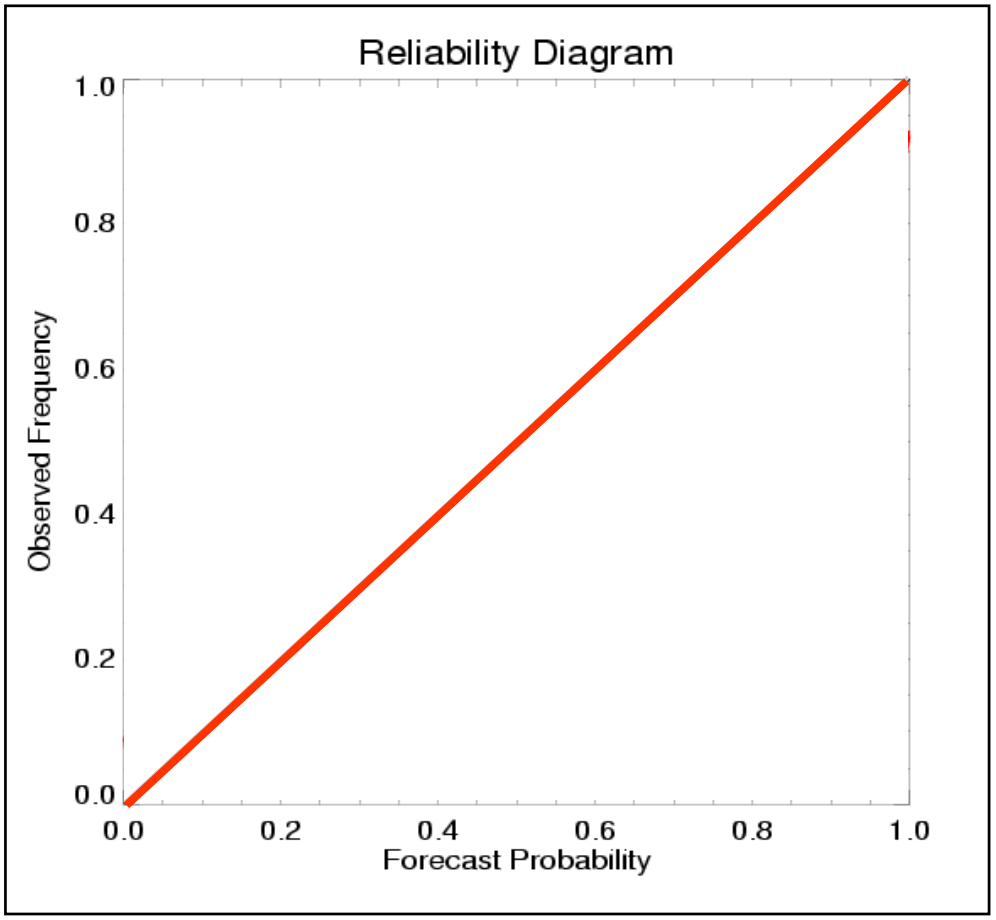


Reliability Diagram

under-confident model

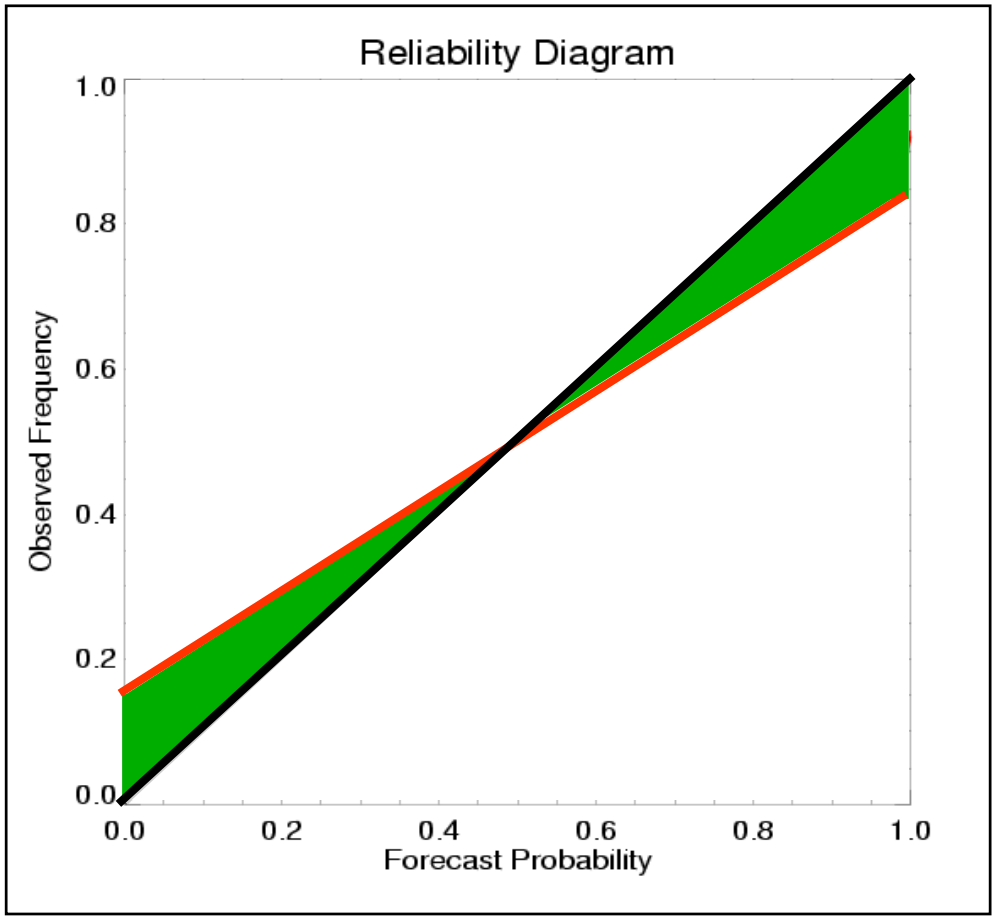


perfect model

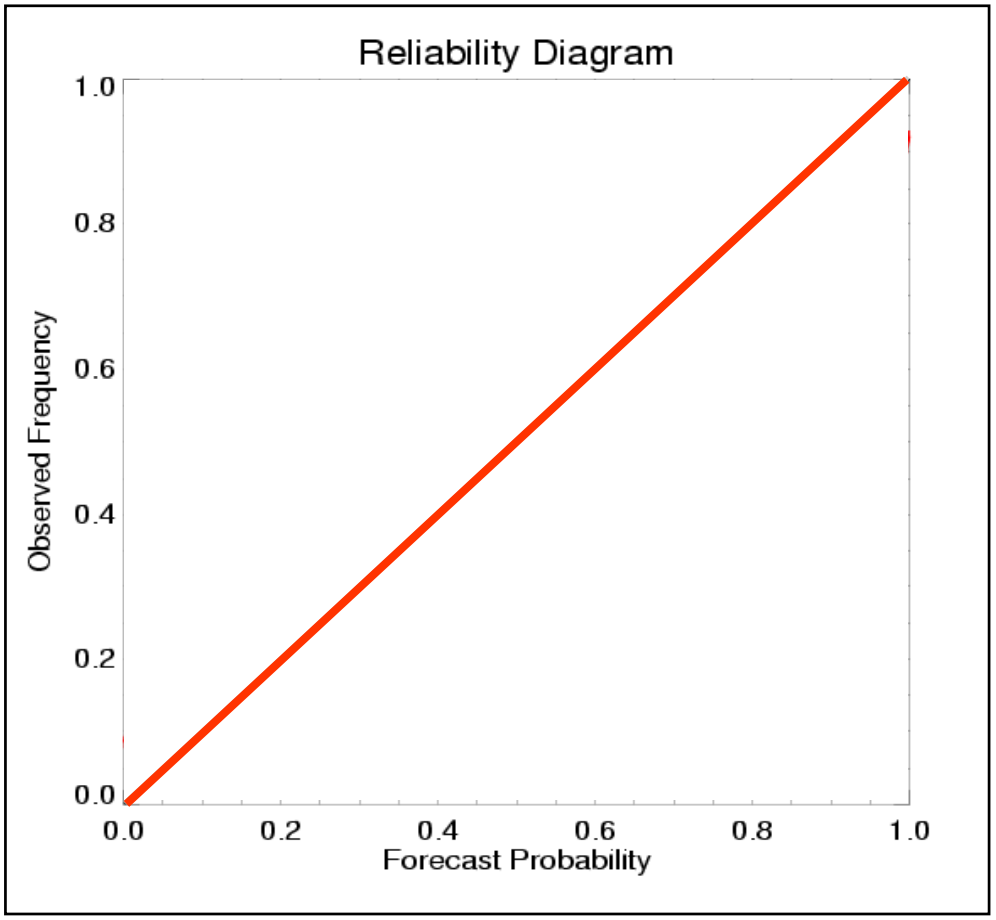


Reliability diagram

Reliability score (the smaller, the better)



imperfect model



perfect model

Components of the Brier Score

$$REL = \frac{1}{N}\sum_{i=1}^{I} n_i (f_i - o_i)^2$$

N = total number of cases

I = number of probability bins

n_i = number of cases in probability bin i

f_i = forecast probability in probability bin I

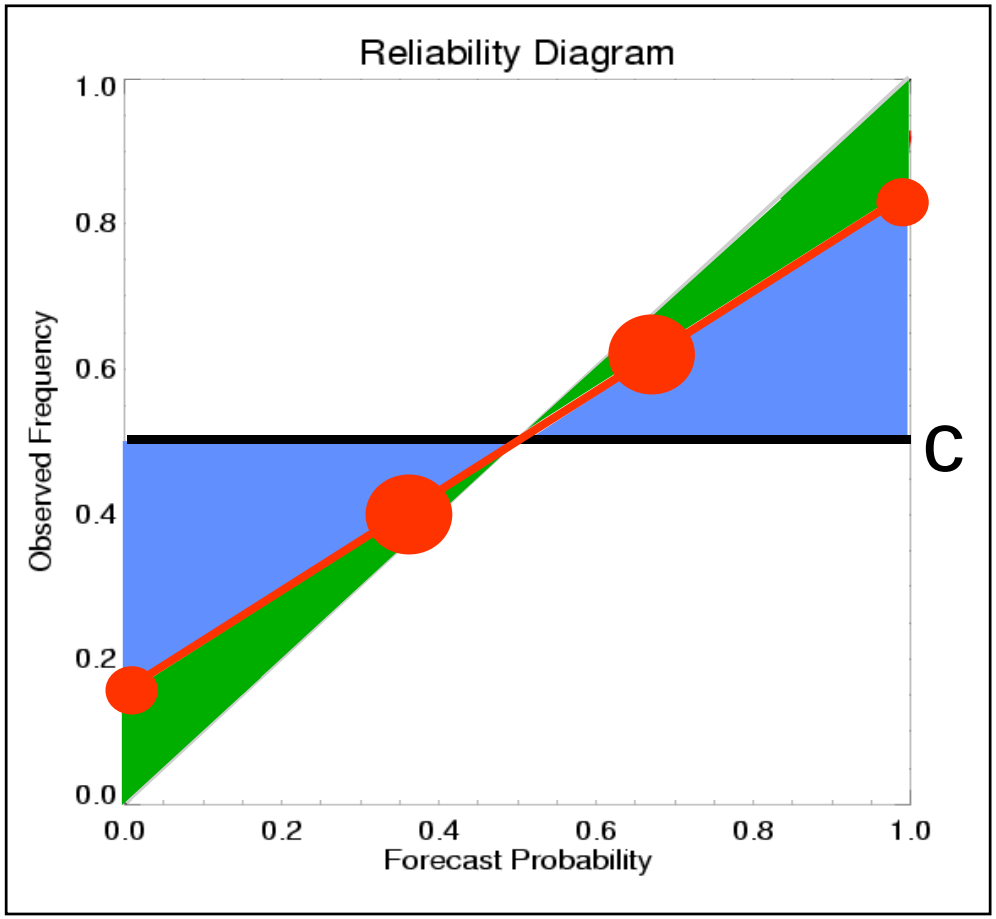
o_i = frequency of event being observed when forecasted with f_i

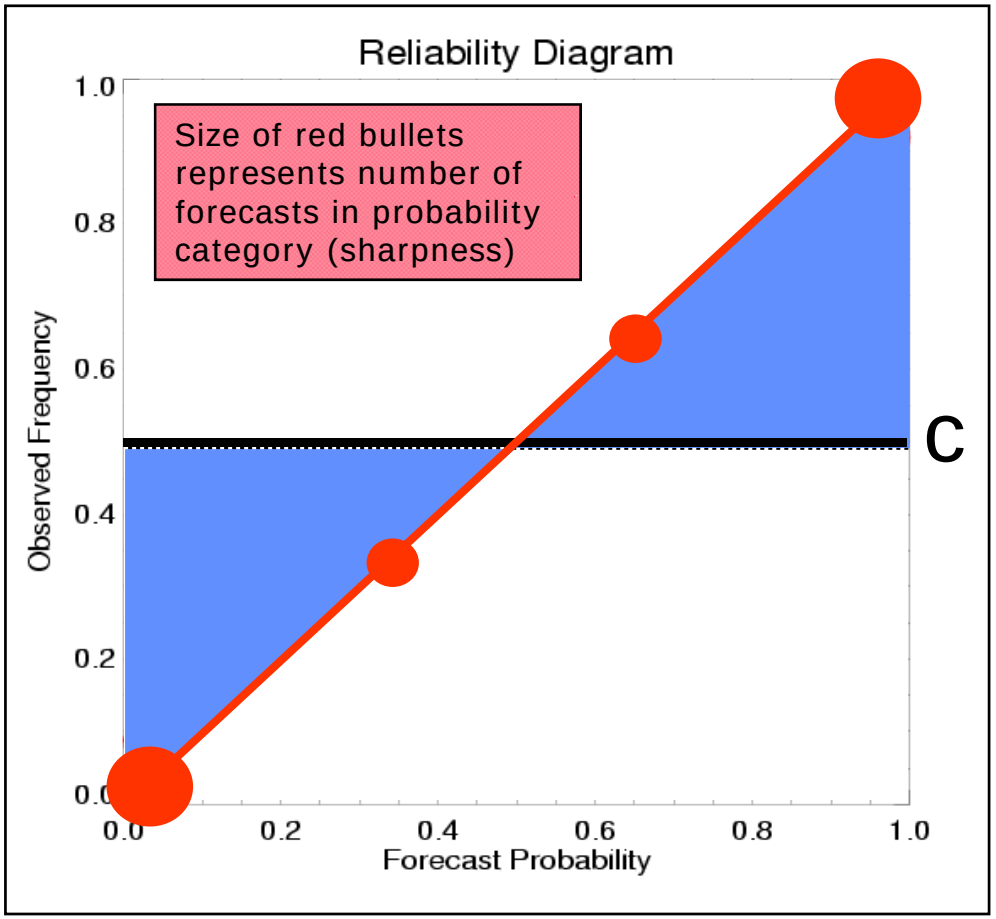
- Reliability: forecast probability vs. observed relative frequencies

Reliability diagram

Reliability score (the smaller, the better)

Resolution score (the bigger, the better)





Poor resolution

Good resolution

Components of the Brier Score

$$REL = \frac{1}{N}\sum_{i=1}^{I} n_i (f_i - o_i)^2$$

$$RES = \frac{1}{N}\sum_{i=1}^{I} n_i (o_i - c\ )^2$$

$$UNC = c(1-c\ )$$

N = total number of cases

I = number of probability bins

n_i = number of cases in probability bin i

f_i = forecast probability in probability bin I

o_i = frequency of event being observed when forecasted with f_i

c = frequency of event being observed in whole sample

- Reliability: forecast probability vs. observed relative frequencies
- Resolution: ability to issue reliable forecasts close to 0% or 100%
- Uncertainty: variance of observations frequency in sample

Brier Score = Reliability – Resolution + Uncertainty

Brier Score

- The Brier score is a measure of the accuracy of probability forecasts

- Considering *N* forecast – observation pairs the BS is defined as:

$$BS = \frac{1}{N}\sum_{n=1}^{N}(p_n - o_n)^2$$

with *p*: forecast probability (fraction of members predicting event)
o: observed outcome (1 if event occurs; 0 if event does not occur)

- BS varies from 0 (perfect deterministic forecasts) to 1 (perfectly wrong!)

- BS corresponds to RMS error for deterministic forecasts

Brier Skill Score

- Skill scores are used to compare the performance of forecasts with that of a reference forecast such as climatology or persistence

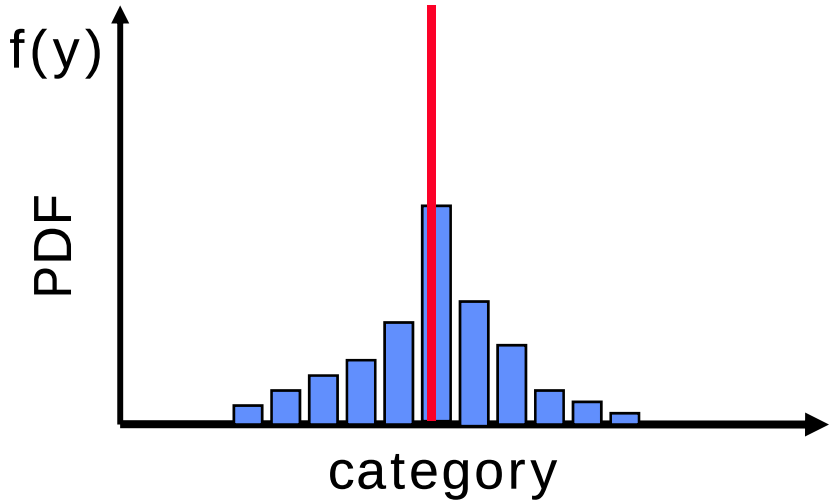
- Constructed so that perfect FC takes value 1 and reference FC = 0

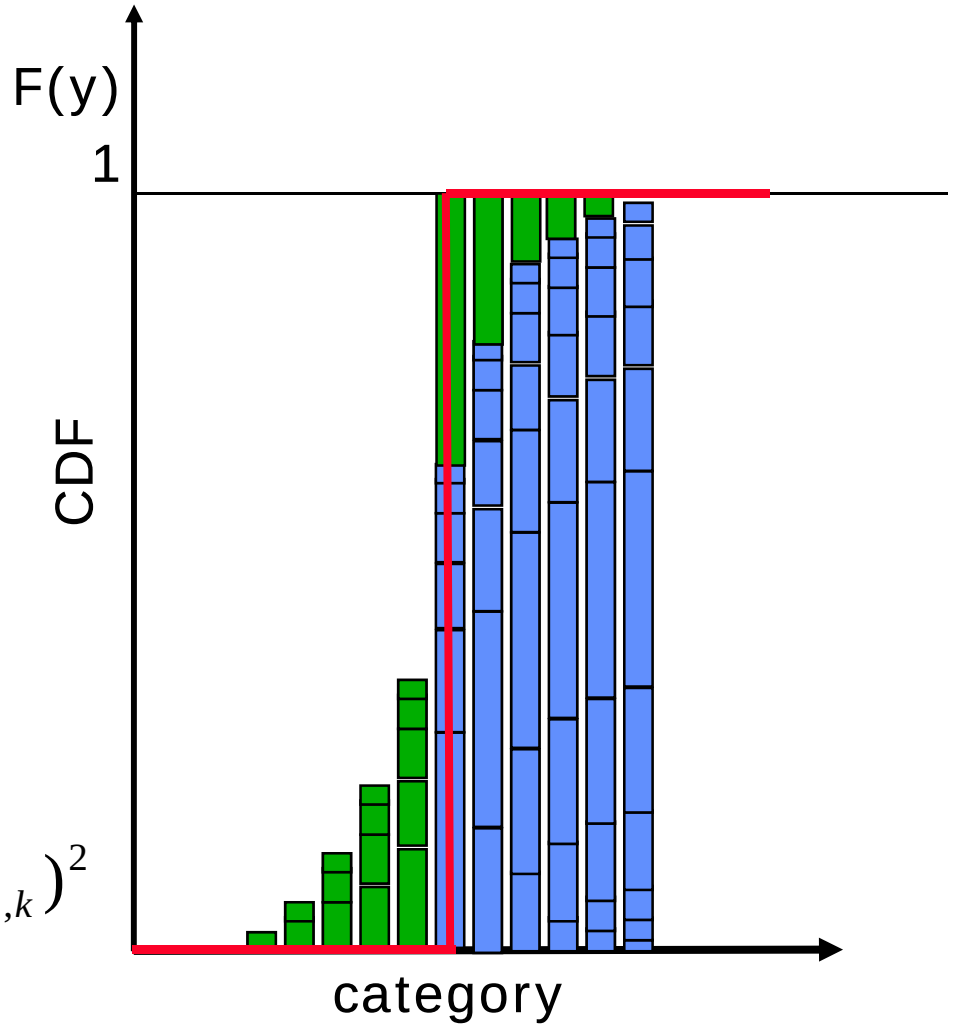
$$\text{Skill score} = \frac{\text{score of current FC} - \text{score for ref FC}}{\text{score for perfect FC} - \text{score for ref FC}}$$

$$BSS = 1 - \frac{BS}{BS_c}$$

- positive (negative) BSS ➢ better (worse) than reference

Ranked Probability Score





$$RPS = \frac{1}{K-1}\sum_{k=1}^{K}(CDF_{FC,k} - CDF_{OBS,k})^2$$

Ranked Probability Score

- Measures the quadratic distance between forecast and verification probabilities for several probability categories *k*
- Emphasizes accuracy by penalizing large errors more than "near misses"
- Rewards sharp forecast if it is accurate
- It is the average Brier score across the range of the variable

$$RPS = \frac{1}{K-1}\sum_{k=1}^{K} BS_k$$

- Ranked Probability Skill Score (RPSS) is a measure for skill relative to a reference forecast

$$RPSS = 1 - \frac{RPS}{RPS_c}$$