

# Ensemble Verification I

Martin Leutbecher



Training Course 2026

# Ensemble Verification I

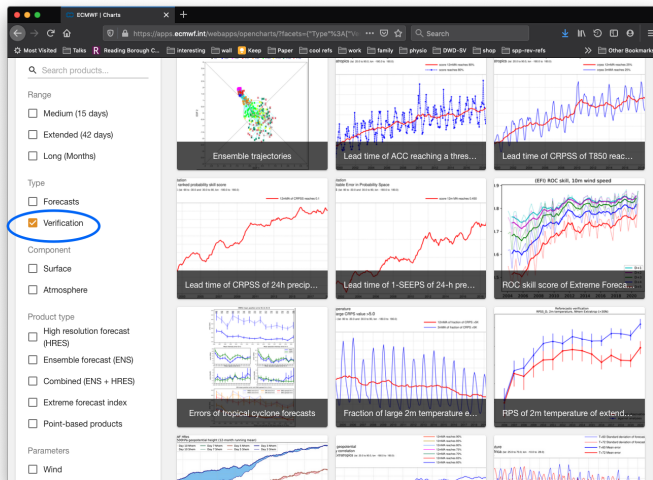
Martin Leutbecher



Training Course 2026

- ① introduction
- ② reliability (statistical consistency)
- ③ dichotomous predictands (yes/no)
  - contingency tables
  - Brier score
  - relative operating characteristic (ROC)
  - logarithmic score
- ④ sensible probabilities:  $p=0$  and  $p=1$ ?

<https://charts.ecmwf.int>



<https://www.ecmwf.int/en/forecasts/quality-our-forecasts>

Assess the quality of a forecast system for

- administrative purposes
  - tool to monitor the system

Assess the quality of a forecast system for

- administrative purposes
  - tool to monitor the system
- scientific/diagnostic purposes
  - Identify strengths and weaknesses of a forecast system
  - Guide the future development of a forecast system
  - Design loss functions for training data driven ensemble models

Assess the quality of a forecast system for

- administrative purposes
  - tool to monitor the system
- scientific/diagnostic purposes
  - Identify strengths and weaknesses of a forecast system
  - Guide the future development of a forecast system
  - Design loss functions for training data driven ensemble models
- economic purposes/ support for decision making
  - Whether a forecast is useful or valuable for a specific user depends on error characteristics but also what other information the user has (eg. climatology) and the particular decision that (s)he needs to make.

Assess the quality of a forecast system for

- administrative purposes
  - tool to monitor the system
- scientific/diagnostic purposes
  - Identify strengths and weaknesses of a forecast system
  - Guide the future development of a forecast system
  - Design loss functions for training data driven ensemble models
- economic purposes/ support for decision making
  - Whether a forecast is useful or valuable for a specific user depends on error characteristics but also what other information the user has (eg. climatology) and the particular decision that (s)he needs to make.
  - An accurate forecast can be of little value (blue desert sky)
  - An inaccurate forecast can be of high value (an intense storm that is predicted but with position error)

- Forecast verification is the investigation of the properties of the **joint distribution of forecasts and observations** (Murphy & Winkler 1987)

- Forecast verification is the investigation of the properties of the **joint distribution of forecasts and observations** (Murphy & Winkler 1987)
- Scalar aspects (attributes) of the forecast quality include:
  - accuracy (e.g. mean absolute error, mean squared error, threat score)
  - bias
  - reliability
  - resolution
  - discrimination
  - sharpness (property of forecast only, e.g. ensemble spread)

- Forecast verification is the investigation of the properties of the **joint distribution of forecasts and observations** (Murphy & Winkler 1987)
- Scalar aspects (attributes) of the forecast quality include:
  - accuracy (e.g. mean absolute error, mean squared error, threat score)
  - bias
  - reliability
  - resolution
  - discrimination
  - sharpness (property of forecast only, e.g. ensemble spread)
- Forecast skill: relative accuracy of one forecast system with respect to a reference forecast (e.g. climatology)
- Forecasts are verified with observations, analyses, or some other estimate of the outcome/true state

# Concepts (II)

Examples of scores for single forecasts

sample of  $N$  forecast-observation pairs  $(x_j, y_j)$ :

- root mean square error  $\left( \frac{1}{N} \sum_{j=1}^N (x_j - y_j)^2 \right)^{1/2}$
- mean absolute error  $\frac{1}{N} \sum_{j=1}^N |x_j - y_j|$
- mean error  $\frac{1}{N} \sum_{j=1}^N (x_j - y_j)$
- anomaly correlation coefficient

sample of  $N$  forecast-observation pairs  $(x_j, y_j)$ :

- root mean square error  $\left( \frac{1}{N} \sum_{j=1}^N (x_j - y_j)^2 \right)^{1/2}$
- mean absolute error  $\frac{1}{N} \sum_{j=1}^N |x_j - y_j|$
- mean error  $\frac{1}{N} \sum_{j=1}^N (x_j - y_j)$
- anomaly correlation coefficient
- scores for dichotomous events (e.g. rain/no rain)
  - Peirce skill score (= Hansen-Kuipers, true skill statistic)
  - Gilbert skill score (Equitable threat score)
  - frequency bias
- All of these scores can be applied to the ensemble mean.

- The ensemble predicted rain with a probability of 10%.
- It did rain on the day
- Is this a good forecasts?
  - Yes
  - No
  - I don't know

- The ensemble predicted rain with a probability of 10%.
- It did rain on the day
- Is this a good forecasts?
  - Yes
  - No
  - I don't know

For probabilistic forecast, the prediction (an ensemble or a probability distribution) and the observation (a value) are different objects. The distribution is not known more precisely after the verifying observation becomes available.

- Are the true values (or observations) statistically indistinguishable from the members of the ensemble?

- Are the true values (or observations) statistically indistinguishable from the members of the ensemble?
- Measures to assess reliability
  - bias
  - “spread” versus “error”
  - rank histogram
  - reliability diagram (for dichotomous (binary) prediction, e.g. rain/no rain or 0/1)

- Are the true values (or observations) statistically indistinguishable from the members of the ensemble?
- Measures to assess reliability
  - bias
  - “spread” versus “error”
  - rank histogram
  - reliability diagram (for dichotomous (binary) prediction, e.g. rain/no rain or 0/1)
- Reliability alone does not imply skill. The climatological distribution is perfectly reliable for a stationary climate.

- Consider ensemble variance ( “spread” ) for an  $M$ -member ensemble

$$\frac{1}{M} \sum_{j=1}^M (x_j - \bar{x})^2$$

and the squared error of the ensemble mean

$$(\bar{x} - y)^2$$

- Average the two quantities for many locations and/or start times.
- The averaged quantities have to match for a reliable ensemble (within sampling uncertainty).

- Consider ensemble variance (“spread”) for an  $M$ -member ensemble

$$\frac{1}{M} \sum_{j=1}^M (x_j - \bar{x})^2$$

and the squared error of the ensemble mean

$$(\bar{x} - y)^2$$

- Average the two quantities for many locations and/or start times.
- The averaged quantities have to match for a reliable ensemble (within sampling uncertainty).
- Finite ensemble size can be corrected for in the estimation of the error of the ensemble mean and the ensemble variance.
- Cave:** Even in a perfect ensemble, the correlation of ensemble spread and rms error is not 1.

# Examples of spread and error

## ECMWF EPS — 500 hPa geopotential, JJA 2017

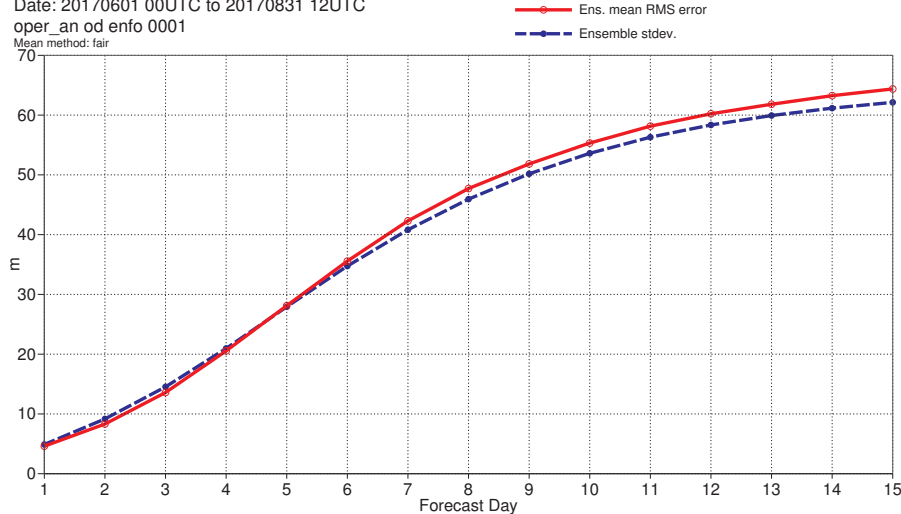
500hPa geopotential

NHem Extratropics (lat 20.0 to 90.0, lon -180.0 to 180.0)

Date: 20170601 00UTC to 20170831 12UTC

oper\_an od enfo 0001

Mean method: fair



# Examples of spread and error

## ECMWF EPS — mean sea level pressure, DJF 2018

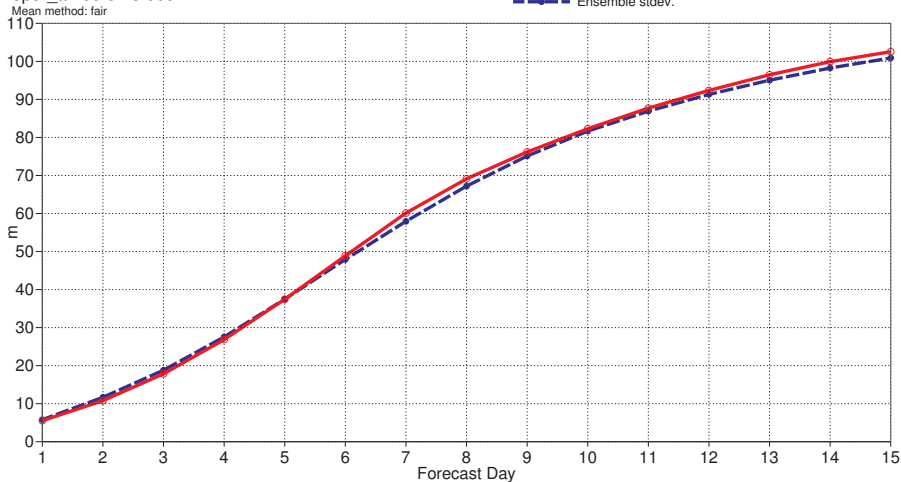
500hPa geopotential

NHem Extratropics (lat 20.0 to 90.0, lon -180.0 to 180.0)

Date: 20171201 00UTC to 20180228 12UTC

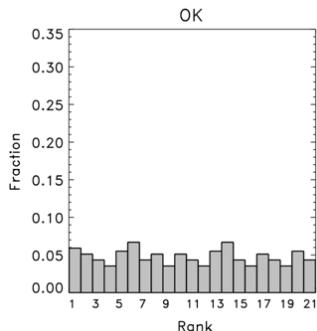
oper\_an od enfo 0001

Mean method: fair

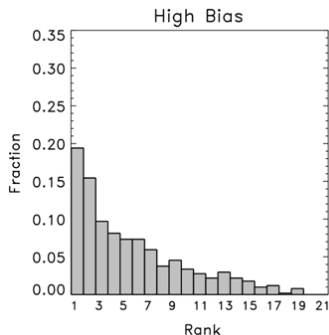


- Are the ensemble members statistically indistinguishable from the verification data?
- Determine where **observation** lies with respect to the ensemble members:

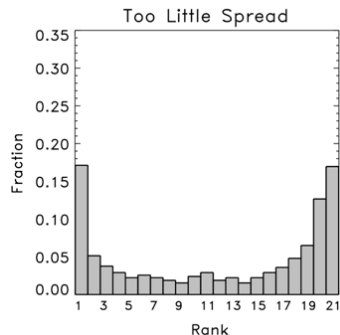




OBS is indistinguishable from any other ensemble member



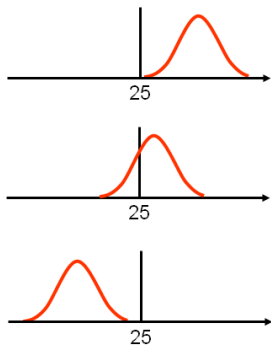
OBS is too often below the ensemble members (biased forecast)



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)

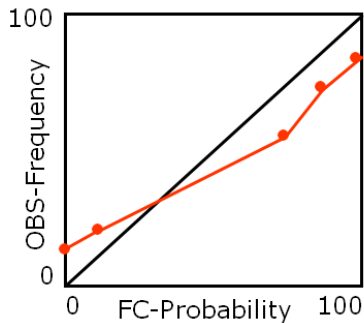
- Consider the probabilistic prediction of the event that the temperature exceeds  $25^{\circ}\text{C}$ .
- Hypothetical verification sample of 30 start dates and 2200 grid points = 66000 forecasts.
- How often was the event ( $T > 25^{\circ}\text{C}$ ) predicted with probability  $p$ ?



| FC Prob. | # FC | OBS-Frequency<br>(perfect model) | OBS-Frequency<br>(imperfect model) |
|----------|------|----------------------------------|------------------------------------|
| 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%)                          |

# Dichotomous predictands

## Reliability diagram

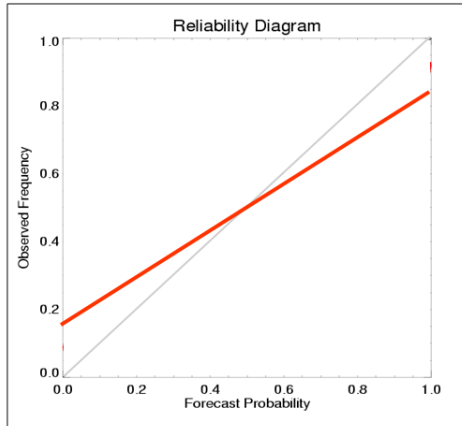


| FC Prob. | # FC | OBS-Frequency<br>(perfect model) | OBS-Frequency<br>(imperfect model) |
|----------|------|----------------------------------|------------------------------------|
| 100%     | 8000 | 8000 (100%)                      | 7200 (90%)                         |
| 90%      | 5000 | 4500 ( 90%)                      | 4000 (80%)                         |
| 80%      | 4500 | 3600 ( 80%)                      | 3000 (66%)                         |
| ....     | .... | ....                             | ....                               |
| ....     | .... | ....                             | ....                               |
| ....     | .... | ....                             | ....                               |
| 10%      | 5500 | 550 ( 10%)                       | 800 (15%)                          |
| 0%       | 7000 | 0 ( 0%)                          | 700 (10%)                          |

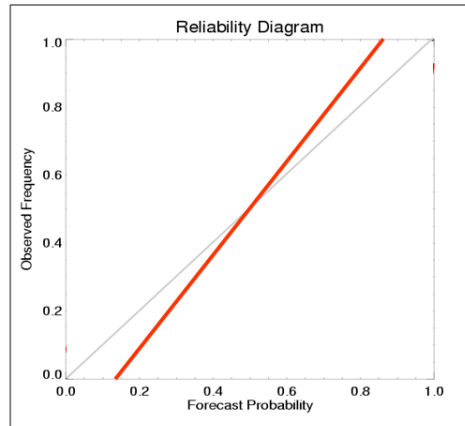
# Over- and under-confidence

Reliability diagram

over-confident model



under-confident model



- Extended contingency tables
- Scores
  - Brier score (reliability and resolution)
  - Logarithmic score (reliability and resolution)
  - Relative Operating Characteristic (discrimination)

- Consider an event  $e$  (e.g.  $T > 25^\circ \text{ C}$ )
- The joint distribution of forecasts and observations can be condensed in a  $2 \times 2$  contingency table:

| $e$ predicted | $e$ observed |                        |
|---------------|--------------|------------------------|
|               | Yes          | No                     |
| Yes           | hits $a$     | false alarms $b$       |
| No            | misses $c$   | correct rejections $d$ |

- hit rate  $H = \frac{a}{a+c}$
- false alarm rate  $F = \frac{b}{b+d}$
- $N = a + b + c + d$  sample size

# (Extended) contingency table

ensemble

The joint distribution of forecasts and observations for a  $M$ -member ensemble can be summarized in a  $(M + 1) \times 2$  contingency table  $\mathbf{T}$

| $e$ pred. by<br>$m_e$ members | $e$ observed |                   |
|-------------------------------|--------------|-------------------|
|                               | Yes          | No                |
| $M$                           | $n_M$        | $\tilde{n}_M$     |
| $M - 1$                       | $n_{M-1}$    | $\tilde{n}_{M-1}$ |
| $\dots$                       | $\dots$      | $\dots$           |
| $j$                           | $n_j$        | $\tilde{n}_j$     |
| $\dots$                       | $\dots$      | $\dots$           |
| 1                             | $n_1$        | $\tilde{n}_1$     |
| 0                             | $n_0$        | $\tilde{n}_0$     |

The joint distribution of forecasts and observations for a  $M$ -member ensemble can be summarized in a  $(M + 1) \times 2$  contingency table  $\mathbf{T}$

$$\text{sample size } N = \sum_{j=0}^M n_j + \sum_{j=0}^M \tilde{n}_j$$

Each row corresponds to a probability value, e.g.  $p = j/M \rightarrow$

| $e$ pred. by<br>$m_e$ members | $e$ observed |                   |
|-------------------------------|--------------|-------------------|
|                               | Yes          | No                |
| $M$                           | $n_M$        | $\tilde{n}_M$     |
| $M - 1$                       | $n_{M-1}$    | $\tilde{n}_{M-1}$ |
| $\dots$                       | $\dots$      | $\dots$           |
| $j$                           | $n_j$        | $\tilde{n}_j$     |
| $\dots$                       | $\dots$      | $\dots$           |
| $1$                           | $n_1$        | $\tilde{n}_1$     |
| $0$                           | $n_0$        | $\tilde{n}_0$     |

The joint distribution of forecasts and observations for a  $M$ -member ensemble can be summarized in a  $(M + 1) \times 2$  contingency table  $\mathbf{T}$

$$\text{sample size } N = \sum_{j=0}^M n_j + \sum_{j=0}^M \tilde{n}_j$$

Each row corresponds to a probability value, e.g.  $p = j/M \rightarrow$

| $e$ pred. by<br>$m_e$ members | $e$ observed<br>Yes | No                |
|-------------------------------|---------------------|-------------------|
| $M$                           | $n_M$               | $\tilde{n}_M$     |
| $M - 1$                       | $n_{M-1}$           | $\tilde{n}_{M-1}$ |
| $\dots$                       | $\dots$             | $\dots$           |
| $j$                           | $n_j$               | $\tilde{n}_j$     |
| $\dots$                       | $\dots$             | $\dots$           |
| 1                             | $n_1$               | $\tilde{n}_1$     |
| 0                             | $n_0$               | $\tilde{n}_0$     |

Contingency tables are additive:

$$\mathbf{T}(\text{sample1} \cup \text{sample2}) = \mathbf{T}(\text{sample1}) + \mathbf{T}(\text{sample2})$$

$$\text{BS} = \frac{1}{N} \sum_{k=1}^N (p_k - o_k)^2$$

- $p_k$  is the predicted probability of the  $k$ -th forecast and  $o_k = 1$  (0) if the event occurred (did not occur)
- The Brier score BS is the **mean squared error** of the probability forecast.

$$\text{BS} = \frac{1}{N} \sum_{k=1}^N (p_k - o_k)^2$$

- $p_k$  is the predicted probability of the  $k$ -th forecast and  $o_k = 1$  (0) if the event occurred (did not occur)
- The Brier score BS is the **mean squared error** of the probability forecast.
- The BS can be decomposed in three components that measure
  - reliability
  - resolution
  - uncertainty

stratify sample in terms of the rows  $j$  in the contingency table

Reliability: deviation of observed relative frequency from forecasted probability

$$\text{REL} = \frac{1}{N} \sum_{j=0}^M \ell_j (\bar{o}_j - p_j)^2$$

$N$  total number of cases

$M$  number of probability bins  $-1$

$p_j = j/M$  probability in bin  $j$

$\ell_j = n_j + \tilde{n}_j$  number of cases in bin  $j$

$\bar{o}_j = n_j/\ell_j$  frequency of event occurring when forecasted with probability  $p_j$

stratify sample in terms of the rows  $j$  in the contingency table

Reliability: deviation of observed relative frequency from forecasted probability

$$\text{REL} = \frac{1}{N} \sum_{j=0}^M \ell_j (\bar{o}_j - p_j)^2$$

Resolution: ability of forecast to identify periods in which observed frequencies differ from average

$$\text{RES} = \frac{1}{N} \sum_{j=0}^M \ell_j (\bar{o}_j - \bar{o})^2$$

- $N$  total number of cases
- $M$  number of probability bins  $-1$
- $p_j = j/M$  probability in bin  $j$
- $\ell_j = n_j + \tilde{n}_j$  number of cases in bin  $j$
- $\bar{o}_j = n_j/\ell_j$  frequency of event occurring when forecasted with probability  $p_j$
- $\bar{o}$  event frequency in whole sample

stratify sample in terms of the rows  $j$  in the contingency table

Reliability: deviation of observed relative frequency from forecasted probability

$$\text{REL} = \frac{1}{N} \sum_{j=0}^M \ell_j (\bar{o}_j - p_j)^2$$

Resolution: ability of forecast to identify periods in which observed frequencies differ from average

$$\text{RES} = \frac{1}{N} \sum_{j=0}^M \ell_j (\bar{o}_j - \bar{o})^2$$

Uncertainty: Variance of obs. (0/1) in sample

$$\text{UNC} = \bar{o}(1 - \bar{o})$$

- $N$  total number of cases
- $M$  number of probability bins  $-1$
- $p_j$   $= j/M$  probability in bin  $j$
- $\ell_j$   $= n_j + \tilde{n}_j$  number of cases in bin  $j$
- $\bar{o}_j$   $= n_j/\ell_j$  frequency of event occurring when forecasted with probability  $p_j$
- $\bar{o}$  event frequency in whole sample

- Skill scores are used to compare the performance of forecasts with that of a reference forecast (e.g. climatological distribution)
- They are defined so that the perfect forecast has a skill score of 1 and the reference forecast has the skill score of 0

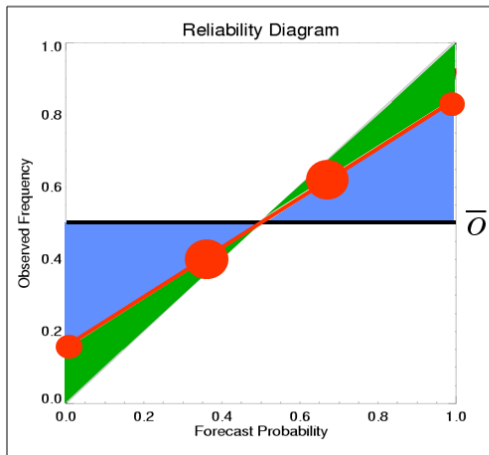
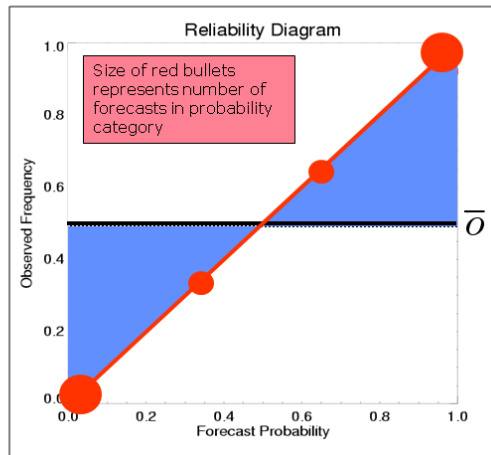
$$\text{skill score} = \frac{\text{actual fc} - \text{ref}}{\text{perfect fc} - \text{ref}}$$

- BS for perfect forecast is 0  $\Rightarrow$

$$\text{BSS} = 1 - \frac{\text{BS}}{\text{BS}_{\text{ref}}}$$

- positive (negative) BSS  $\Rightarrow$  forecast is better (worse) than the reference forecast

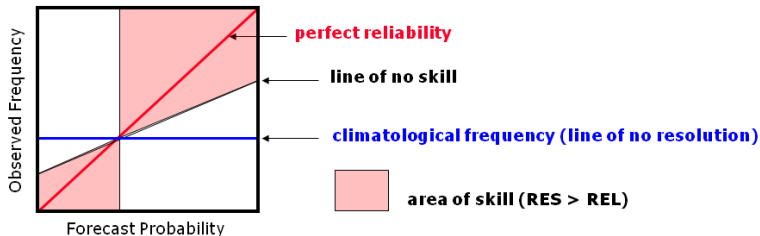
- Reliability score (the smaller, the better)
- Resolution score (the bigger, the better)



# Positive contribution to skill

diagnosed from the attributes diagram

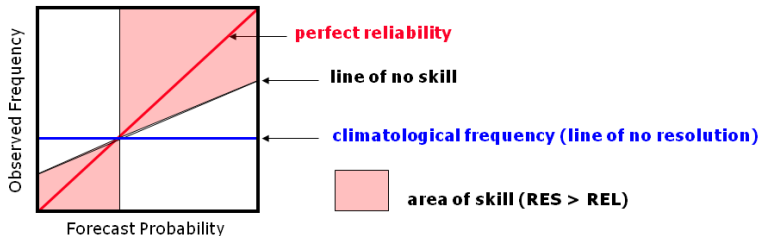
$$\begin{aligned} BSS &= 1 - \frac{BS}{BS_c} \\ &= 1 - \frac{REL - RES + UNC}{UNC} = \frac{RES - REL}{UNC} \end{aligned}$$



# Positive contribution to skill

diagnosed from the attributes diagram

$$BSS = 1 - \frac{BS}{BS_c}$$
$$= 1 - \frac{REL - RES + UNC}{UNC} = \frac{RES - REL}{UNC}$$

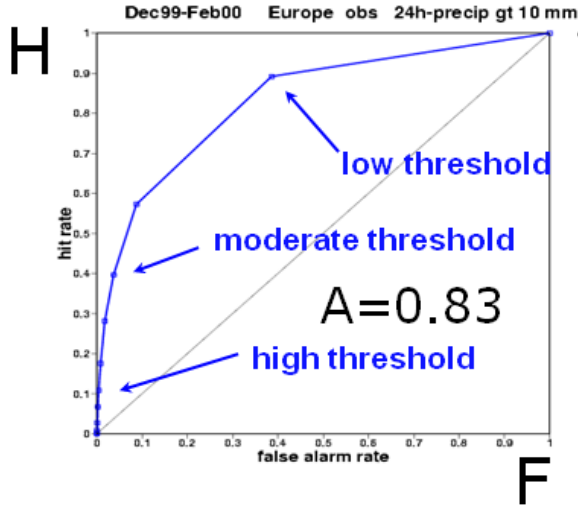


Cave: Using sample climatology as reference can lead to fictitious skill

- Until now, we asked:  
What is the distribution of observations  $o$  if the forecast system predicts an event to occur with probability  $p$ ?
- To measure the ability of a forecast system to *discriminate* between occurrence and non-occurrence of an event, we have to ask:  
What is the distribution of forecast probabilities when the event occurred and what is the distribution when it did not occur?

- Until now, we asked:  
What is the distribution of observations  $o$  if the forecast system predicts an event to occur with probability  $p$ ?
- To measure the ability of a forecast system to *discriminate* between occurrence and non-occurrence of an event, we have to ask:  
What is the distribution of forecast probabilities when the event occurred and what is the distribution when it did not occur?
- For evaluation purposes, let us predict the event when the probability exceeds a threshold  $p_i$ .
- For any probability threshold  $p_i$ , compute the hit rate  $H_i = \frac{a}{a+c}$  and the false alarm rate  $F_i = \frac{b}{b+d}$
- The *relative operating characteristic* (ROC, also referred to as receiver operating characteristic) is the diagram that shows  $H$  versus  $F$  for all probability thresholds.

# Relative Operating Characteristic



- random forecast (independent of observed event) on diagonal
- summary measure: area under the ROC  $\in [0.5, 1]$

- also known as ignorance score (Good 1952, Roulston and Smith 2002)

$$\text{LS} = -\frac{1}{N} \sum_{k=1}^N [o_k \log p_k + (1 - o_k) \log(1 - p_k)]$$

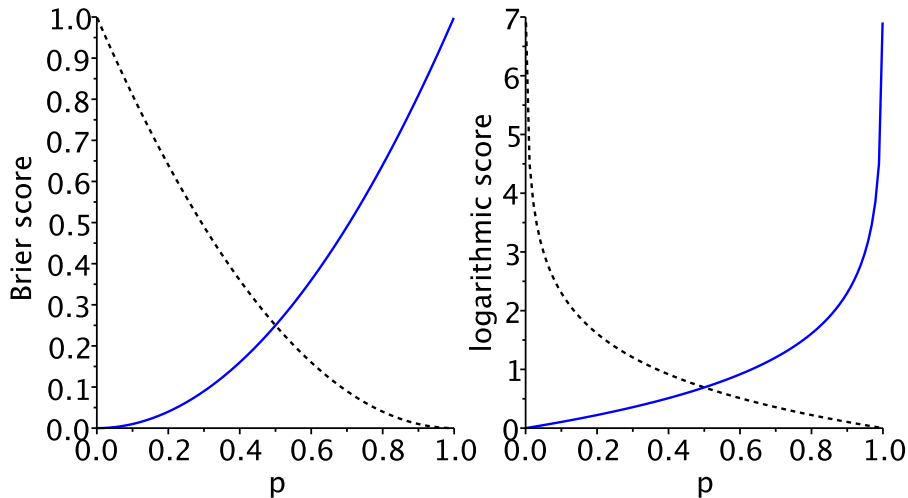
- also known as ignorance score (Good 1952, Roulston and Smith 2002)

$$LS = -\frac{1}{N} \sum_{k=1}^N [o_k \log p_k + (1 - o_k) \log(1 - p_k)]$$

- The score ranges between 0 and  $\infty$ . The latter happens if the predicted probability is zero and the event occurs (or if  $p = 1$  and the event does not occur).
- The ignorance score is more sensitive to the cases with probability close to 0 and close to 1 than the Brier score.

## Brier score versus logarithmic score

event occurs (dotted), event does not occur (solid)  
 $(p - 1)^2$  and  $p^2$        $-\log(p)$  and  $-\log(1 - p)$

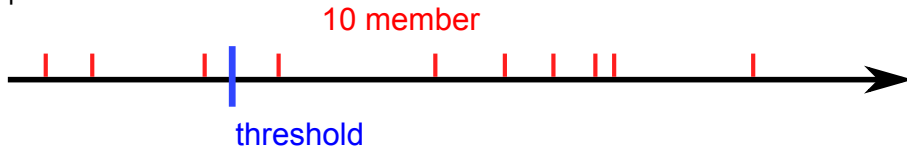


- Never forecast  $p = 0$  or  $p = 1$  unless you are really certain!
- If the true probability is not equal to zero (or one), there will still be cases when no member (or all members) predict(s) the event.  
Sampling uncertainty!

- Never forecast  $p = 0$  or  $p = 1$  unless you are really certain!
- If the true probability is not equal to zero (or one), there will still be cases when no member (or all members) predict(s) the event.

Sampling uncertainty!

- Wilks proposed to estimate cumulative probabilities using Tukey's plotting positions



- When  $n$  members of an  $M$ -member ensemble have a value less than the threshold  $\theta$ , the probability to not exceed  $\theta$  is set to

$$p^{(T)}(n) = \frac{n + 2/3}{M + 4/3}$$

- Consider for example  $M = 10$ :

|   |      |      |      |      |      |      |      |      |      |      |      |
|---|------|------|------|------|------|------|------|------|------|------|------|
| n | 0    | 1    | 2    | 3    | 4    | 5    | 6    | 7    | 8    | 9    | 10   |
| p | 0.06 | 0.15 | 0.24 | 0.32 | 0.41 | 0.50 | 0.59 | 0.68 | 0.76 | 0.85 | 0.94 |