Ensemble Verification I

Martin Leutbecher

ECMWF

Training Course 2019

Ensemble Verification I

Martin Leutbecher

ECMWF

Training Course 2019

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?

Objectives of verification (. . . evaluation and diagnostics)

Assess the quality of a forecast system for

- administrative purposes
    - tool to monitor the system

Objectives of verification (. . . evaluation and diagnostics)

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

Objectives of verification (. . . evaluation and diagnostics)

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
- 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.

Objectives of verification (. . . evaluation and diagnostics)

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
- 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)

Objectives of verification (. . . evaluation and diagnostics)

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
- 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)
  - The actual forecast value may differ from the potential forecast value (availability of relevant fc information, user's constraints: economic, time limits, lack of training, etc.)

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

Concepts

- 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)

Concepts

Forecast attributes and forecast skill

- 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)
- More generally: observations $\rightarrow$ estimates of the true state (e.g. also analyses)

Concepts (II)

Examples of scores for single forecasts

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

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
- 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.

Concepts (III)

Probabilistic forecasts and ensemble forecasts

- 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

Concepts (III)

Probabilistic forecasts and ensemble forecasts

- 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.

Statistical consistency and reliability

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

Statistical consistency and reliability

- 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)

definitions and examples . . .

Statistical consistency and reliability

- 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)

  definitions and examples . . .
- Reliability alone does not imply skill. The climatological distribution is perfectly reliable for a stationary climate.

Reliability of the ensemble spread

- Consider ensemble variance ("spread") for an M-member ensemble

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

and the squared error of the ensemble mean

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

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

Reliability of the ensemble spread

- Consider ensemble variance ("spread") for an M-member ensemble

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

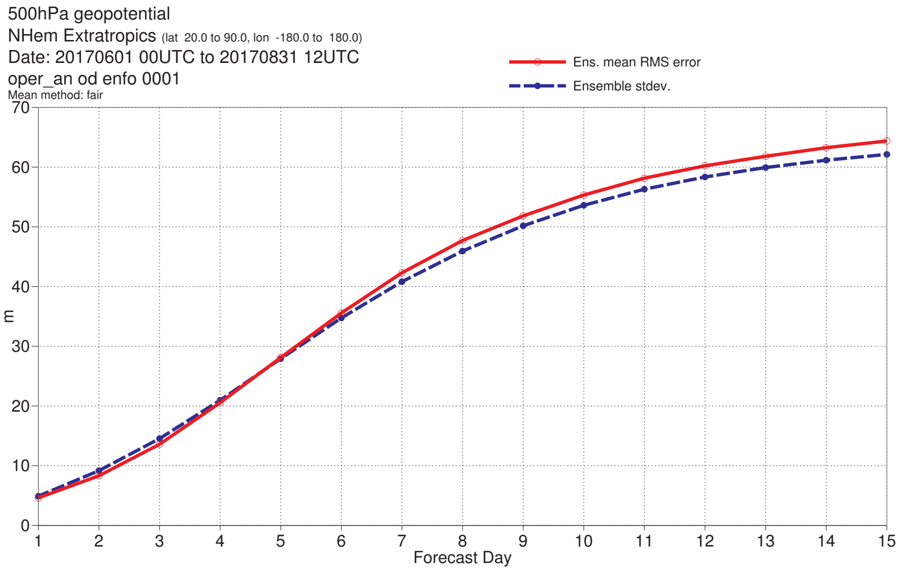
and the squared error of the ensemble mean

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

- 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.

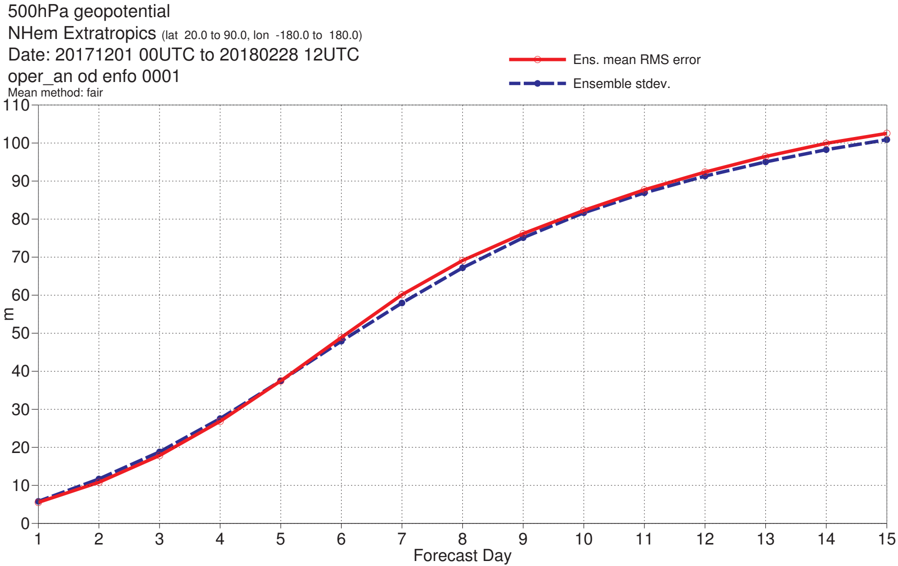
Examples of spread and error

ECMWF EPS — 500 hPa geopotential, JJA 2017



Examples of spread and error

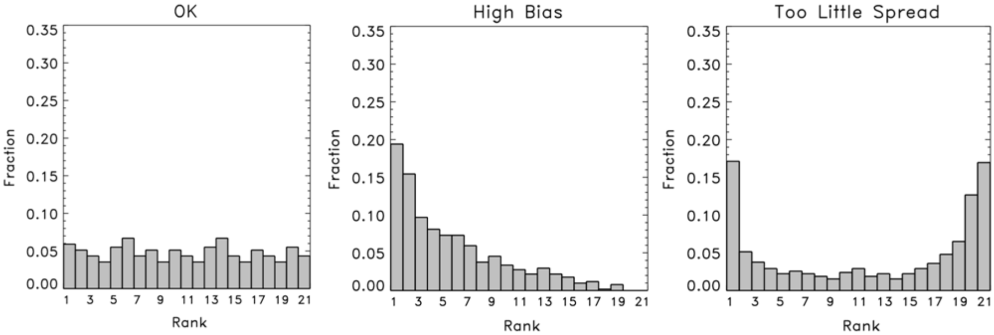
ECMWF EPS — mean sea level pressure, DJF 2018



Rank Histogram

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

Rank Histogram



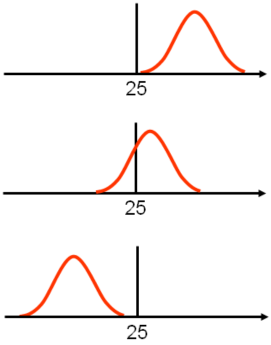
OBS is indistinguishable from any other ensemble member

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)

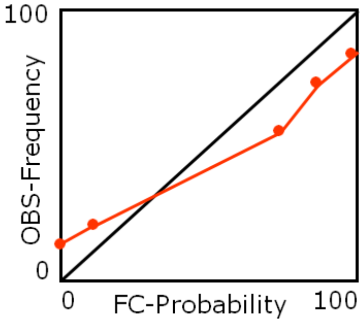
- Consider the probabilistic prediction of the event that the temperature exceeds 25° C.
- Hypothetical verification sample of 30 start dates and 2200 grid points = 66000 forecasts.
- How often was the event ($T > 25°$ C) predicted with probability p?



FC Prob.	# FC	OBS-Frequency (perfect model)	OBS-Frequency (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

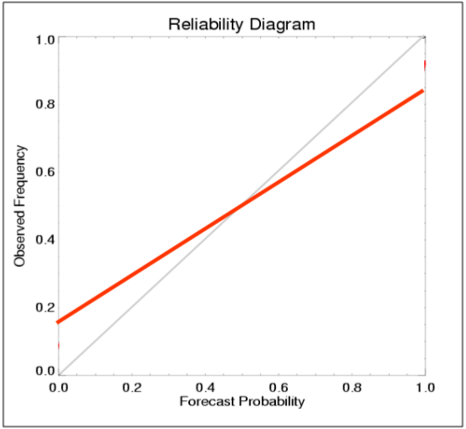


FC Prob.	# FC	OBS-Frequency (perfect model)	OBS-Frequency (imperfect model)
100%	8000	8000 (100%)	7200 (90%)
90%	5000	4500 (90%)	4000 (80%)
80%	4500	3600 (80%)	3000 (66%)
....	....	....	....
....	....	....	....
....	....	....	....
10%	5500	550 (10%)	800 (15%)
0%	7000	0 (0%)	700 (10%)

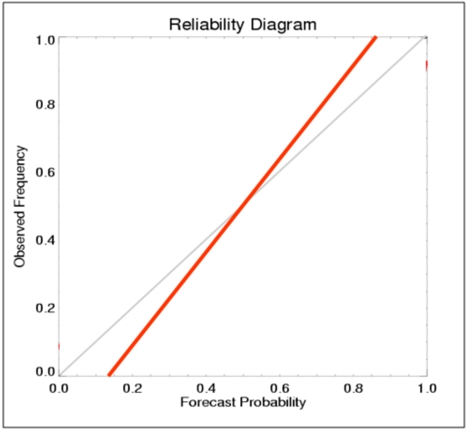
Over- and under-confidence

Reliability diagram

over-confident model



under-confident model



Scores for dichotomous predictions

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

Contingency table

single forecast

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

	e observed	
e predicted	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

ECMWF

(Extended) contingency table

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	e observed	
m_e members	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	$\tilde{n}_0$

(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 $\mathbf{T}$

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

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

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

(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 $\mathbf{T}$

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. $p = j/M \quad \longrightarrow$

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

Contingency tables are additive:

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

Brier score

definition and decomposition

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

- 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.

Brier score

definition and decomposition

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

- 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

Brier score components

BS=REL−RES+UNC

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

Reliability: deviation of observed relative frequency from forecasted probability

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

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

Brier score components

BS=REL−RES+UNC

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

Reliability: deviation of observed relative frequency from forecasted probability

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

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

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

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

Brier score components

BS=REL−RES+UNC

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

Reliability: deviation of observed relative frequency from forecasted probability

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

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

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

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

$$\mathrm{UNC} = \overline{o}(1 - \overline{o})$$

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

Brier Skill Score

- 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 $\Rightarrow$

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

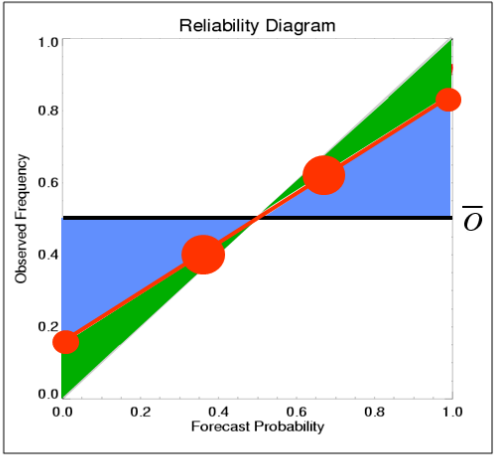
- positive (negative) BSS $\Rightarrow$ forecast is better (worse) than the reference forecast

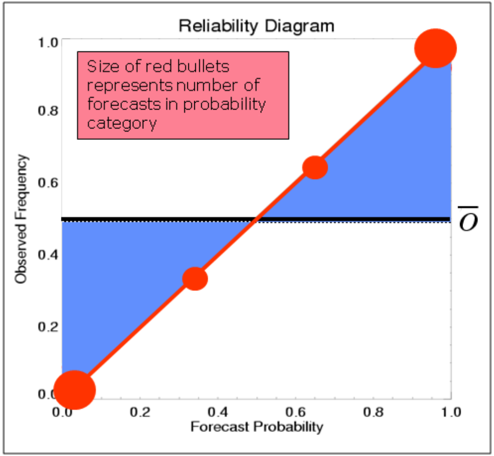
Brier score

Attributes diagram

Reliability score (the smaller, the better)

Resolution score (the bigger, the better)



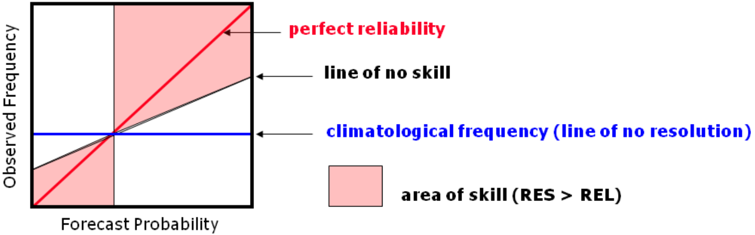


Positive contribution to skill

diagnosed from the attributes diagram

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

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

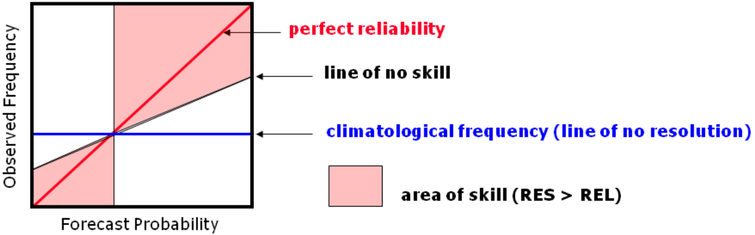


Positive contribution to skill

diagnosed from the attributes diagram

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

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

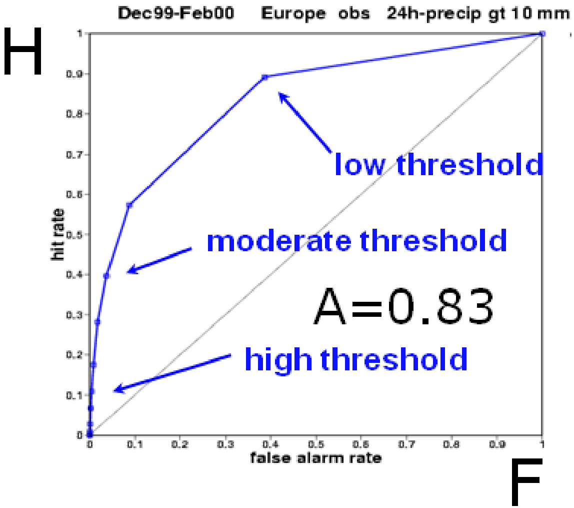


Cave: Using sample climatology as reference can lead to ficticious skill

Discrimination and ROC

- until now, we looked at question:
  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, one has to ask:
  What distributions of probabilities have been predicted when the event occurred and when it did not occur?
- For any probability threshold p_i one can then determine 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 $\in [0.5, 1]$

Logarithmic score

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

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

Logarithmic score

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

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

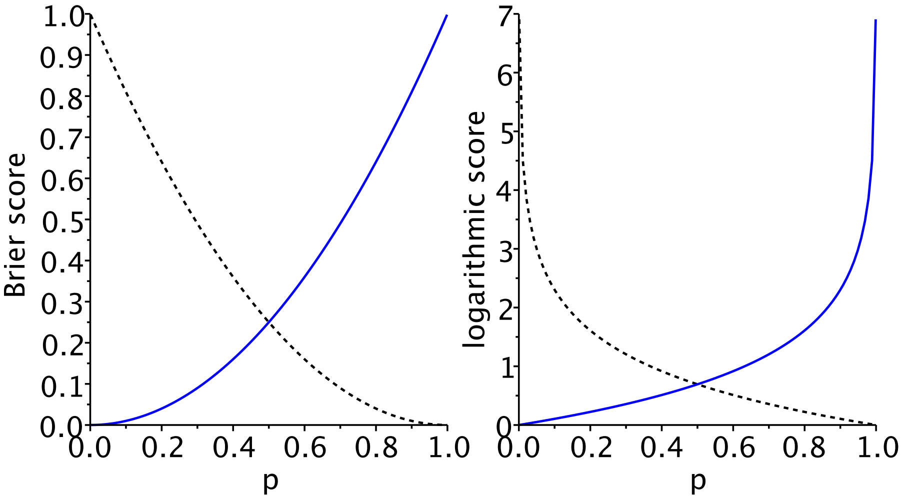
- The score ranges between 0 and ∞. The latter happens if the predicted probability is zero and the event occurs
  (or if $p = 1$ and the event does not occur).
- The ignorance score is more sensitive to the cases with probability close to 0 and close to 1 than the Brier score.

ECMWF

Brier score versus logarithmic score

event occurs (dotted), event does not occur (solid)

$(p-1)^2$ and p^2 $\qquad$ $-\log(p)$ and $-\log(1-p)$

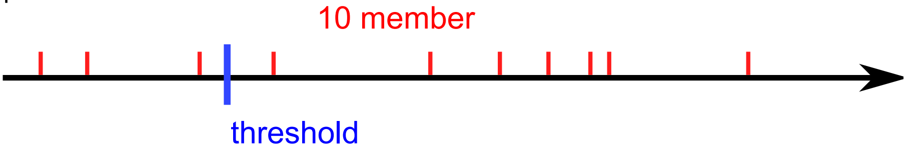


Sensible probabilities

- 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!

Sensible probabilities

- 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 θ, the probability to not exceed θ 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