Seasonal Forecast Verification
Forecast Attributes

JP. Céron
and contribution from S. Mason (IRI)
jpceron.wmo@gmail.com

WMO OMM

World Meteorological Organization
Organisation météorologique mondiale

lectures

1 *Introduction*

2 **Forecast Attributes**

3 *ROC & Reliability - Exercise*

4 *Significance & Robustness*

5

Forecast Attributes

As a general rule in forecast verification, there is no single verification measure or score that can assess all the characteristics of a forecast whatever probabilistic or deterministic.

Forecast quality is multi-faceted, and more than one verification measure is needed to assess the different aspects of forecast quality.

The various aspects of verification which should be adressed are called forecast attributes

Forecast Attributes

Accuracy

- **Definition** :
  - This is the **degree of agreement of the forecasts with the observations** (or reference data)**.**
- **Measurement** :
  - In the case of probability forecasts, the observations are binary ; 1 if the forecast event occurs and 0 if it doesn’t. For probabilistic forecast, the accuracy can be measured by score such as the Brier Score
  - For deterministic forecast, the accuracy is usually assessed by means of a score such as the Mean Square Error

Both summarize the accuracy of the forecast into a single number over a verification dataset.

Forecast Attributes

Mean Squared Error

$$\text{Mean Squared Error (MSE)} = \frac{\text{total of squared errors}}{\text{number of forecasts}}$$

Brier Score

$$\text{MSE in probability} = \frac{\text{total of squared probability errors}}{\text{number of forecasts}}$$

A measure of the forecast **accuracy** (sensitive to the mean bias).

The bias measurement is important in weather forecast verification, but less so in seasonal forecasting because of calibration.

Forecast Attributes

Skill (or Skill-Score)

- **Definition** :
  - **The accuracy of the forecast with reference to the accuracy of a reference forecast** such as Climatology, persistance or another model.
  - The skill score defines the **percentage improvement** in accuracy over the reference forecast.
- **Measurement** :
  - The most commonly used standard is "climatology", defined either as the frequency with which the event was observed over the verification sample (or the long-term frequency if known), or the climatological mean.
  - The skill is scaled that the reference forecast has a skill of 0 and that you are **better** than the reference if your **skill is > 0**
  - All scores can be transformed into Skill-Scores (e.g. MSSS, BSS, …)

  - The skill is a relative measurement instead an absolute one (like MSE), **you can get « very high » skill if you compare your forecast (even if not very accurate) with a very bad forecast.**

Finley's Tornado Forecasts

A set of tornado forecasts for the U.S. Midwest published in 1884.

	FORECASTS		
OBSERVATIONS	**Tornado**	**No tornado**	***Total***
Tornado	**28**	**23**	**51**
No tornado	**72**	**2680**	**2752**
Total	**100**	**2703**	**2803**

$$\text{Heidke score} = \frac{28+2680}{2803} \times 100 = 96.6$$

Other Forecast : No Tornado Forecast

A better score can be achieved by issuing no forecasts of tornadoes!

	FORECASTS		
OBSERVATIONS	**Tornado**	**No tornado**	***Total***
Tornado	**0**	**51**	**51**
No tornado	**0**	**2752**	**2752**
Total	**0**	**2803**	**2803**

$$\text{Heidke score} = \frac{0+2752}{2803} \times 100 = 98.2$$

Other : always Tornado Forecast

Another strategy: score can be achieved by issuing always forecasts of tornadoes!

	FORECASTS		
OBSERVATIONS	**Tornado**	**No tornado**	***Total***
Tornado	**51**	**0**	**51**
No tornado	**2752**	**0**	**2752**
Total	**2803**	**0**	**2803**

$$\text{Heidke score} = \frac{51+0}{2803} \times 100 \sim 1.8$$

Skill-Score Principle

For a given score (e.g. Heidke Score), one compare the score value got from the evaluated forecasting system to the score value got from a reference forecasting system

$$\text{Heidke Skill Score} = \frac{\text{Correct Forecast - Reference Correct Forecasts}}{\text{Perfect Forecast - Reference Correct Forecasts}} \times 100$$

How many more times was the forecast correct compared to a reference forecast strategy?

$$\boxed{\text{Heidke skill score} = \frac{(28+2680)-2752}{2803\text{-}2752} \times 100 = -86.3}$$

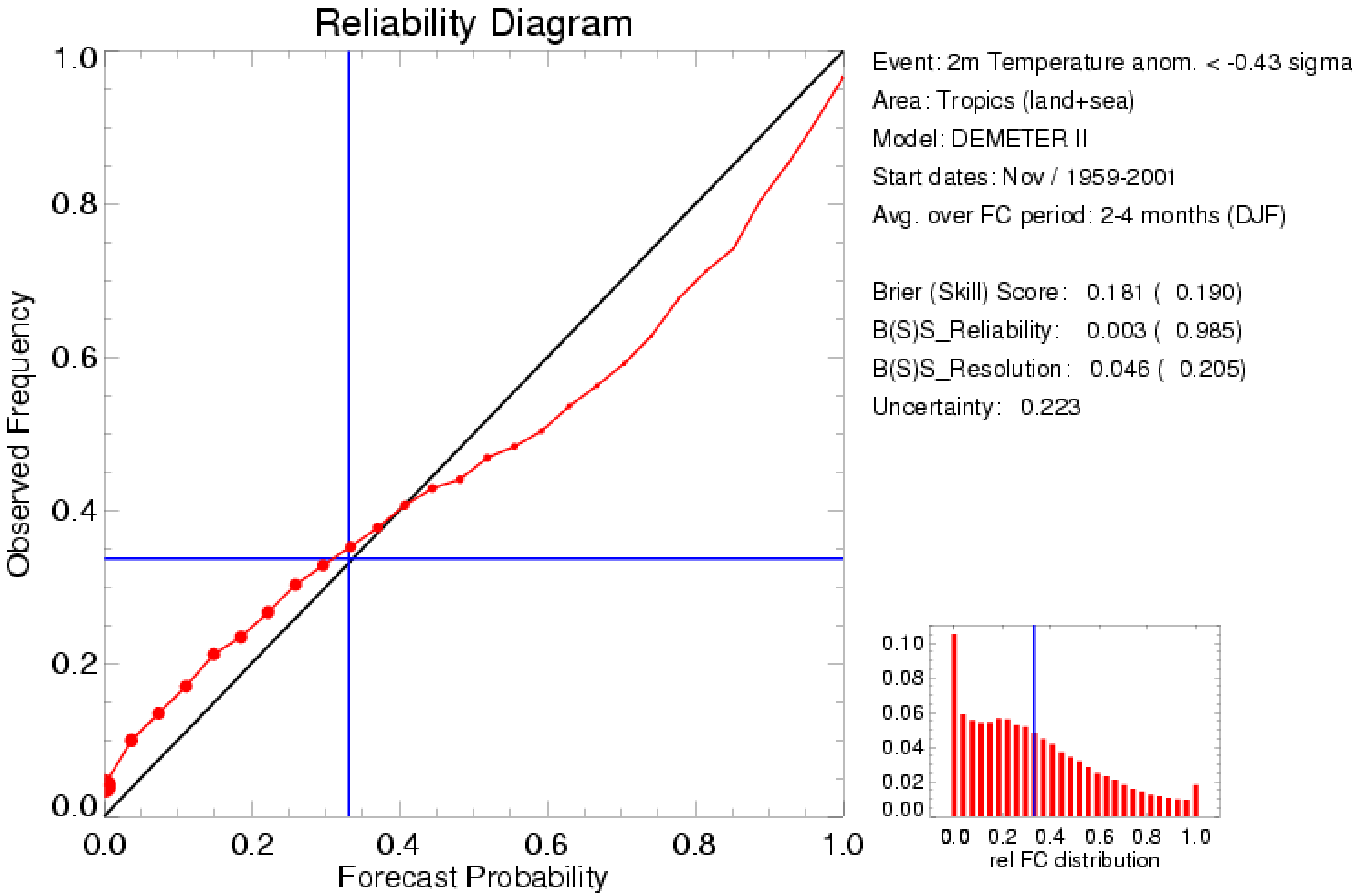
$$\boxed{\text{Heidke skill score} = \frac{(28+2680)-51}{2803\text{-}51} \times 100 = +96.5}$$

Forecast Attributes

Reliability

- **Definition** :
  - For all occasions when a specific probability is forecasted, **the difference between the frequency of occurrence of the event and the forecast probability is low**..
  - For example, if the event occurs around 80% of the time when a probability close to 80% has been forecasted, then this probability forecast should said to be reliable.
- **Measurement** :
  - Reliability is like bias for deterministic forecasts. When the frequency of occurrence is higher (lower) than the forecast probability then there is said to be an under- (over-) forecasting bias in the probability
  - It is investigated using reliability diagrams.
  - ***Reliability cannot be assessed on a single forecast***; a collection of forecasts is needed; **the larger the sample the better it is.**

Reliability Diagram



Event: 2m Temperature anom. < -0.43 sigma
Area: Tropics (land+sea)
Model: DEMETER II
Start dates: Nov / 1959-2001
Avg. over FC period: 2-4 months (DJF)

Brier (Skill) Score: 0.181 (0.190)
B(S)S_Reliability: 0.003 (0.985)
B(S)S_Resolution: 0.046 (0.205)
Uncertainty: 0.223

Reliability diagram for the DJF season over Tropics T2m – Lower Tercile (Below Normal) DEMETER II

Forecast Attributes

Resolution

- **Definition :**
  - **It** refers to the **ability of the forecast system to systematically distinguish between subsets of the sample with different frequencies of occurrence of the event**., .
  - For example, if one compares the observed frequency of occurrence on all those occasions when 20% is forecasted with the frequency of occurrence of the event when 80% is forecasted, there should be a difference. If there is a difference of any kind, then the forecast is said to have resolution.
- **Measurement** :
  - ***Resolution also cannot be assessed on a single forecast.***
  - Resolution for probabilistic forecasts is like Anomaly Correlation Coefficient for deterministic forecasts

Forecast Attributes

Sharpness

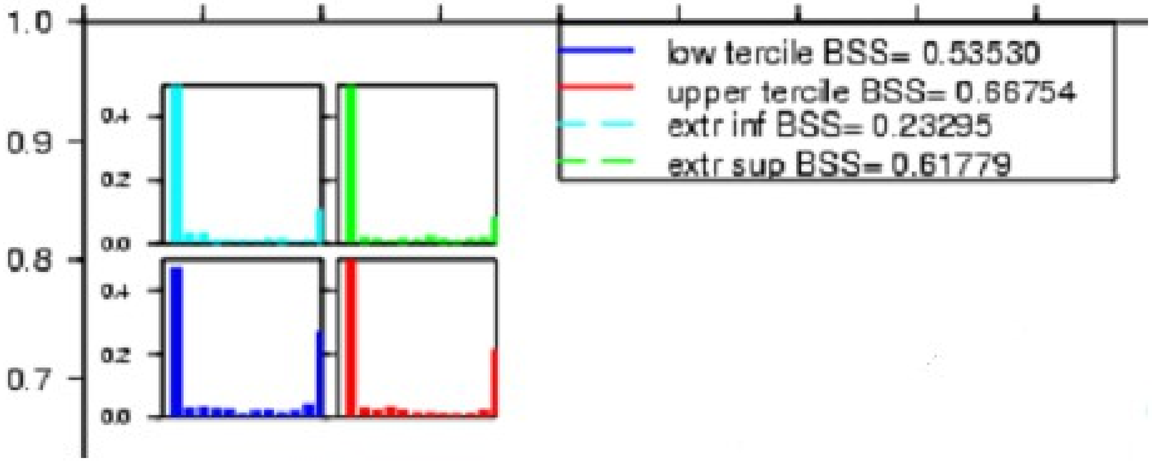
- **Definition** :
  - Sharpness refers only to **the distribution of forecast probabilities** over a verification sample.
  - **This is an attribute of the forecasts only**, unlike the foregoing attributes, All preceeding attribute require a matched set of forecasts and observations to be measured. .
  - For example, if **probabilities near 0 and 1 (100%) are often used** , then the forecast is said « **sharp** ». If most of the forecast probabilities are in a short range of probabilities (e.g. 25 to 40%) then this forecast system would be said "smooth" or « unsharp" .
- **Measurement** :
  - Sharpness can be measured by the variance or the standard deviation of the forecast probabilities.
  - It is investigated using the pdf of the probabilistic forecast

Forecast Attributes

5. Sharpness :

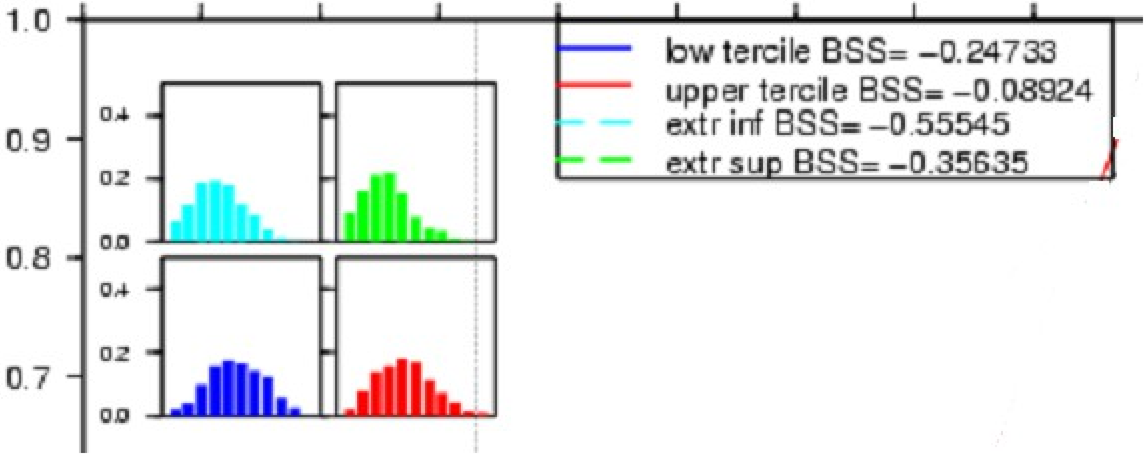
Niño 3.4 SST

DJF – MF system 3



T2m Northern Europe

DJF – MF system 3



Forecast Attributes

Discrimination

- **Definition** :
  - This is the **ability of the forecast system to distinguish between occurrences of the event and non-occurrences** by forecasting a different set of probabilities when the event occurres than when it doesn't. .
- **Measurement** :
  - Discrimination is evaluated by measuring **the difference between the two conditional distributions of forecast probabilities**, the distribution of forecasts when the event occurres and the distribution of forecasts given the non-occurrence of the event.
  - All measurement able to quantify the distance between the 2 conditionnal forecasted probability distributions could be relevant.(e.g. Mahalanobis distance, Khi2 distance, ROC area under the ROC curve, ...)
  - ***Discrimination is similar to resolution, but not the same. Discrimination is in terms of the forecasts conditioned on the observations, while resolution is conditioned on the forecasts.***

Forecast Attributes

Uncertainty

- **Definition** :
  - This attribute is the converse to sharpness, referring to the distribution of the observations only.
  - For an event related to the frequency of occurrence

The observed frequency of an event1 is **5%**

The observed frequency of an event2 is **45%**

The observed frequency of the event3 is **90%**

which case corresponds to the greater uncertainty ?

- **Measurement** :
  - The uncertainty can be measured by **the variance of the observations**. Since the observations are binary, the variance is given by ***p(1-p)*** where *p* is the frequency with which the event occurs in the sample, sometimes called the *sample climatology*.

Verification of probabilistic forecasts

- How do we know if a probabilistic forecast was “correct”?

“A probabilistic forecast can never be wrong!”

As soon as a forecast is expressed probabilistically, all possible outcomes are forecast. However, the forecaster’s level of confidence can be “correct” or “incorrect” = **reliable**.

Is the forecaster **over-** / **under-confident**?

Reliability and Attributes Diagrams

For all forecasts of a given confidence, identify how often the event occurs. If the proportion of times that the event occurs is the same as the forecast probability, the probabilities are **reliable** (or **well calibrated**).

A plot of relative frequency of occurrence against forecast probability will be a diagonal line if the forecasts are **reliable**.

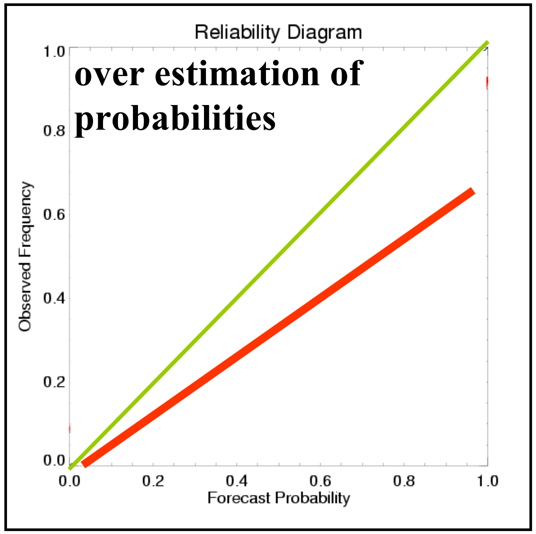
Problem: large number of forecasts required.

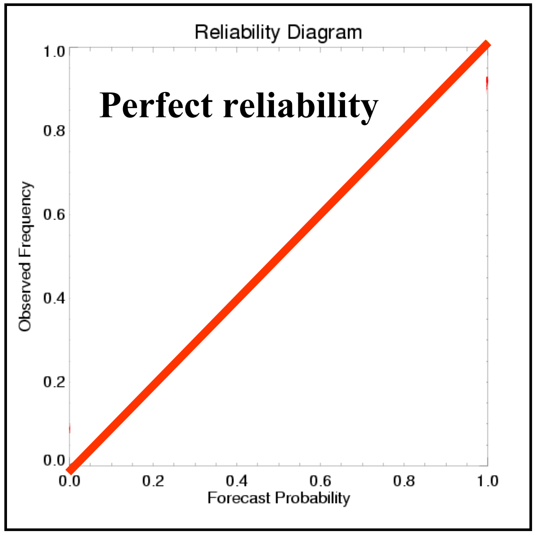
Verification of probabilistic forecasts

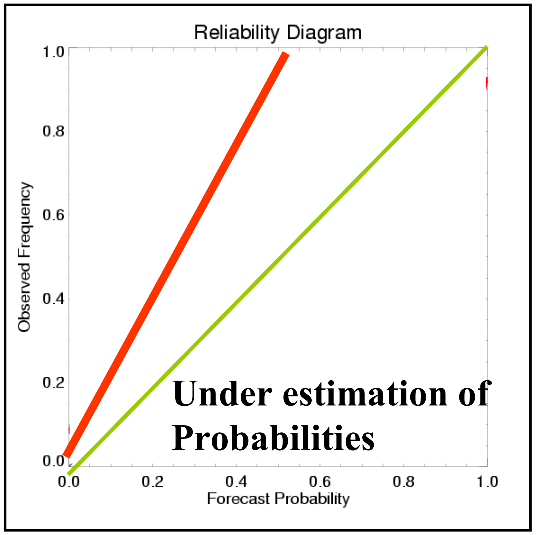
- How do we know if a forecaster is over- / under-confident

Whenever a forecaster says there is a high probability of rain tomorrow, it should rain more frequently than when the forecaster says there is a low probability of rain.

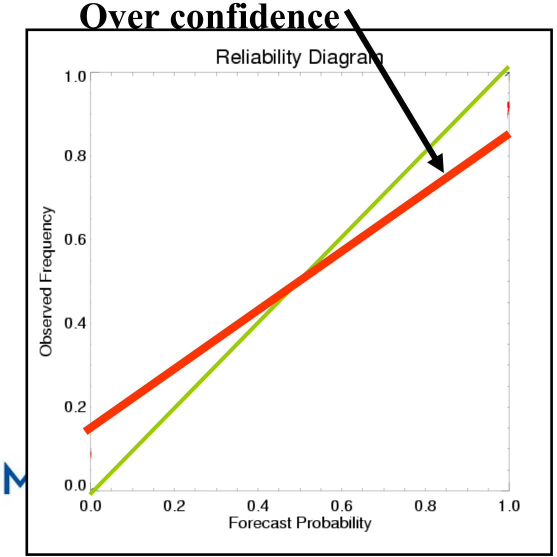
Interpretation of reliability diagrams


over estimation of probabilities

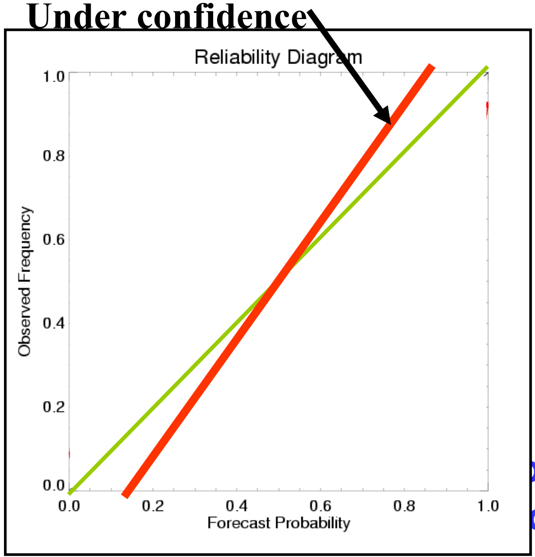

Perfect reliability


Under estimation of Probabilities

Over confidence



Under confidence



Multi faceted scores

Brier score

Measures the mean-squared error of probability forecasts *(equivalent of MSE for deterministic forecast)*.

$$\text{Brier score} = \frac{\text{total of squared probability errors}}{\text{number of forecasts}}$$

If an event was forecast with a probability of 60%, and the event occurred, the probability error is:
60% - 100% = -40%   and BS contribution is 0.16

Multi faceted scores

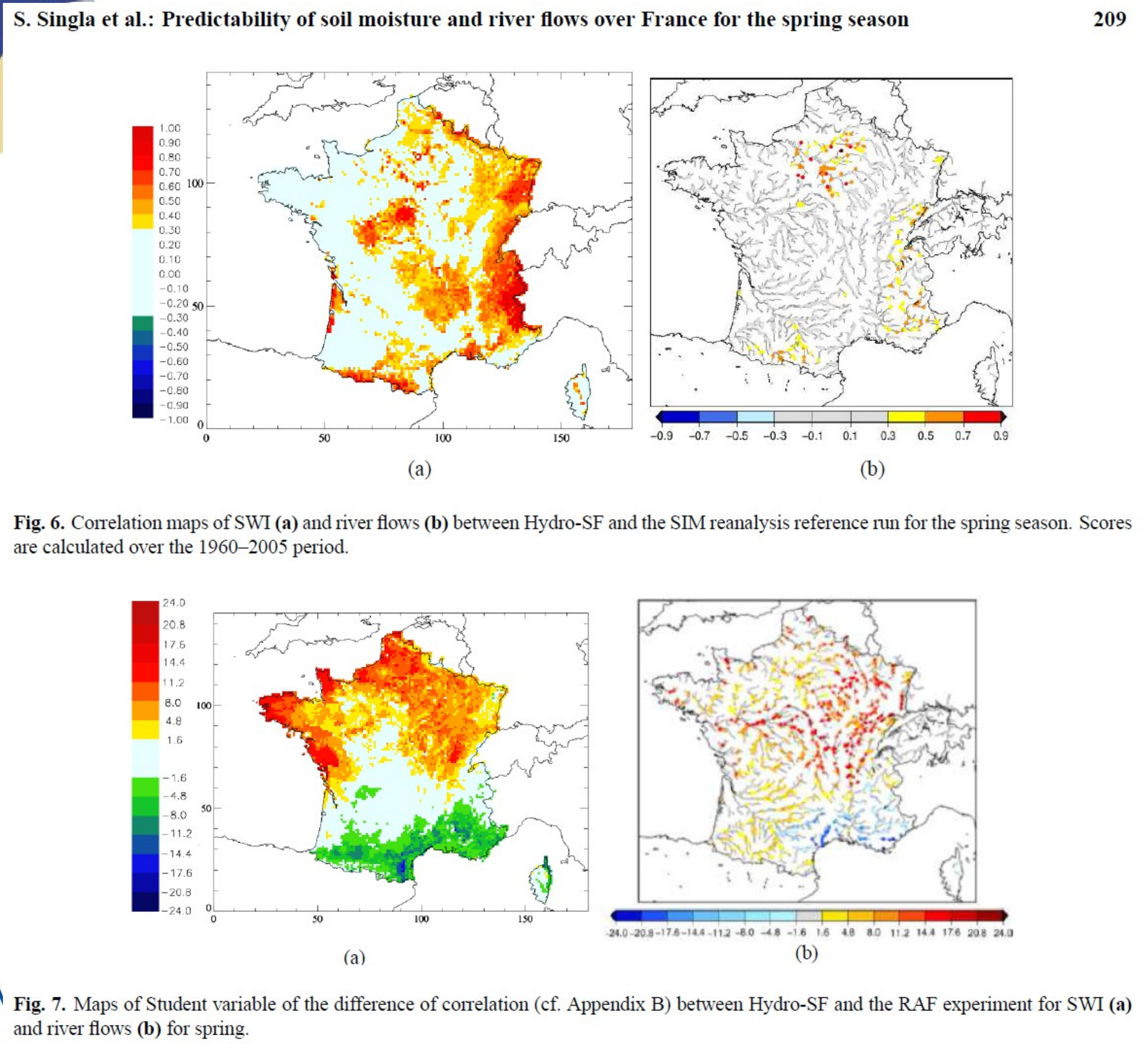
Brier score (*Murphy’s decomposition*)

Brier score = reliability – resolution + uncertainty

Resolution : when the forecast is 60% for dry, is the outcome the same as when the forecast is 10% for dry?

Reliability : when the forecast is 60% for dry, do dry conditions occur 60% of the time?

Uncertainty : what is the climatological probability of dry conditions occurring?



1.00
0.90
0.80
0.70
0.60
0.50
0.40
0.30
0.20
0.10
0.00
−0.10
−0.20
−0.30
−0.40
−0.50
−0.60
−0.70
−0.80
−0.90
−1.00
0
50
100
150
−0.9
−0.7
−0.5
−0.3
−0.1
0.1
0.3
0.5
0.7
0.9
(a)
(b)

Fig. 6. Correlation maps of SWI **(a)** and river flows **(b)** between Hydro-SF and the SIM reanalysis reference run for the spring season. Scores are calculated over the 1960–2005 period.

24.0
20.8
17.6
14.4
11.2
8.0
4.8
1.6
−1.6
−4.8
−8.0
−11.2
−14.4
−17.6
−20.8
−24.0
0
50
100
150
-24.0 -20.8 -17.6 -14.4 -11.2 -8.0 -4.8 -1.6 1.6 4.8 8.0 11.2 14.4 17.6 20.8 24.0
(a)
(b)

Fig. 7. Maps of Student variable of the difference of correlation (cf. Appendix B) between Hydro-SF and the RAF experiment for SWI **(a)** and river flows **(b)** for spring.

W

shop
a

BS

S. Singla et al.: Predictability of soil moisture and river flows over France for the spring season 211

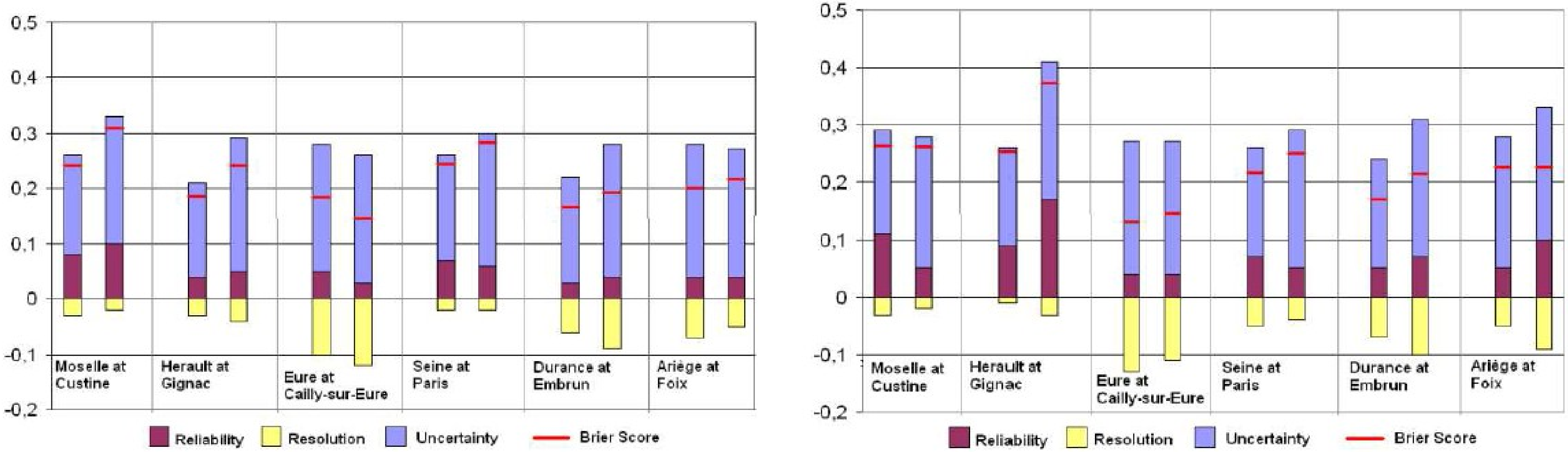


Fig. 10. Histograms of the decomposition of Brier Score (reliability, resolution, uncertainty) (A2) and Brier Score (A1) for river flow forecasts from RAF (left panel) and Hydro-SF (right panel) for Spring over the 1960–2005 period. Graphs show the results from 6 different river catchments for the upper (left bar) and lower (right bar) tercile categories.

WMO OMM

Pre-COF Training Workshop
15-18/11/2016 - Roma

Multi faceted scores

Ranked probability score

The same as the Brier score, but for multiple categories.

The Brier score and the ranked probability score can be expressed as **skill scores** in the same way as for the Heidke (hit) score.

Verification measures for continuous probabilistic forecasts are experimental – there are very few attempts to estimate the full probability distribution of possible outcomes.

Multi faceted scores

$$MSSS = 1 - \frac{MSE_F}{MSE_c}$$

where : MSE=Mean Squared Error

F for Forecasts

c for Climatology

Murphy decomposition :

$$MSSS = \left\{ 2\frac{\mathrm{var}_F}{\mathrm{var}_o} cor_{Fo} - \left(\frac{\mathrm{var}_F}{\mathrm{var}_o} \right)^2 - \left(\frac{\left[mean_F - mean_o \right]^2}{\mathrm{var}_o} \right) + \frac{2n-1}{(n-1)^2} \right\} \left\{ 1 + \frac{2n-1}{(n-1)^2} \right\}$$

Phase error

Amplitude error

Systematic error

avec : **var** = variance

mean = mean

cor = correlation

n = size of sampling

o for observation

Forecast Attributes

ATTRIBUTE	DEFINITION
Accuracy	Average correspondence between individual pairs of observations and forecasts
Skill	Accuracy of forecasts relative to accuracy of forecasts produced by a standard method
Reliability	Correspondence of conditional mean observation and conditioning forecasts, averaged over all forecasts
Resolution	Difference between conditional mean observation and unconditional mean observation, averaged over all forecasts
Sharpness	Variability of forecasts as described by distribution forecasts
Discrimination	Difference between conditional mean forecast and unconditional mean forecast, averaged over all observations
Uncertainty	Variability of observations as described by the distribution of observations

From Murphy - 1993

Forecast Attributes

ATTRIBUTE	RELATED MEASURES
Accuracy	mean absolute error (MAE), mean squarred error (MSE), root mean squared error, Brier score (BS)
Skill	Brier skill score, others in the usual format
Reliability	Reliability component of BS, MAE, MSE of binned data from reliability table
Resolution	Resolution component of BS
Sharpness	Variance of forecasts
Discrimination	Area under ROC, measures of separation of conditional distribution ; MAE, MSE of scatter plot, binned by observation value
Uncertainty	Variance of observations

From Murphy - 1993

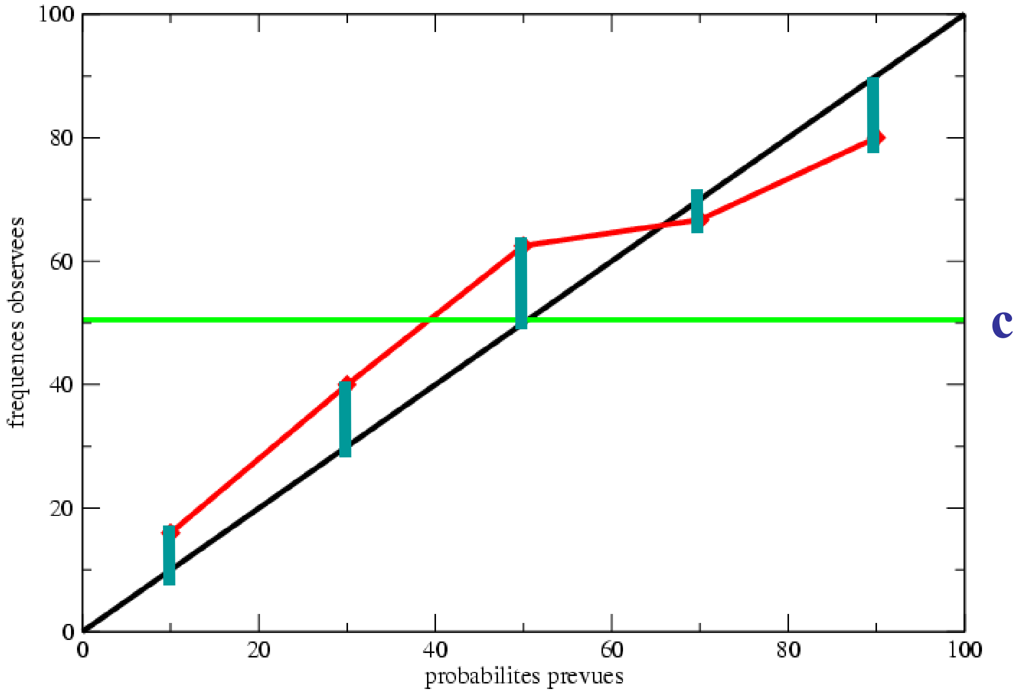
WEATHER CLIMATE WATER
TEMPS CLIMAT EAU

WMO OMM

World Meteorological Organization
Organisation météorologique mondiale

Reliability

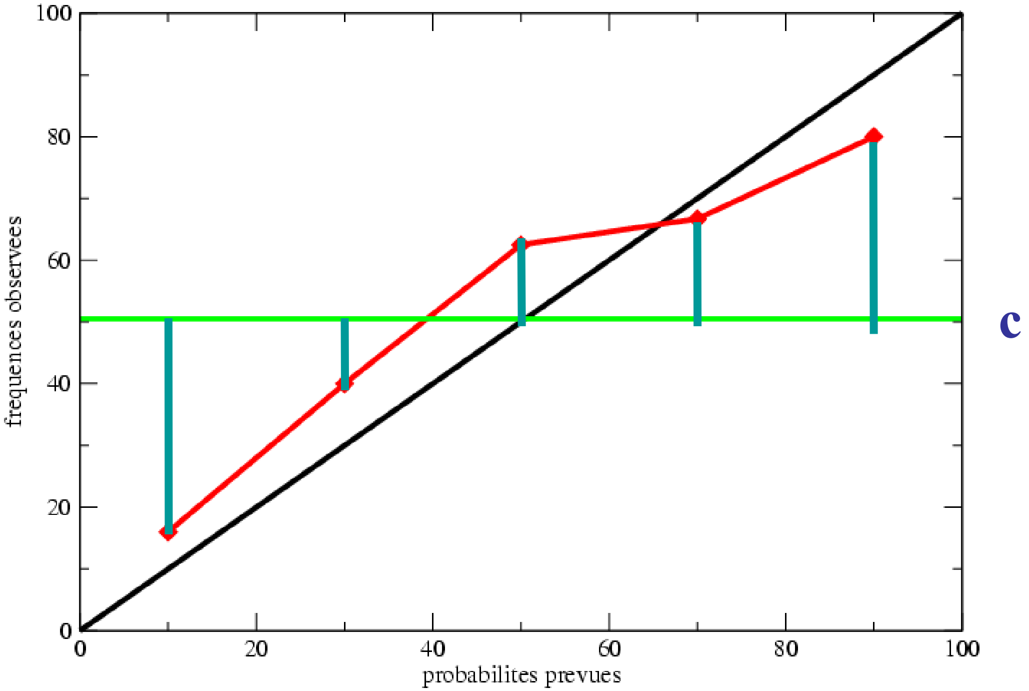
$$reliability = \frac{1}{N} \sum_{i=1}^{I} n_i (p_i - o_i)^2$$



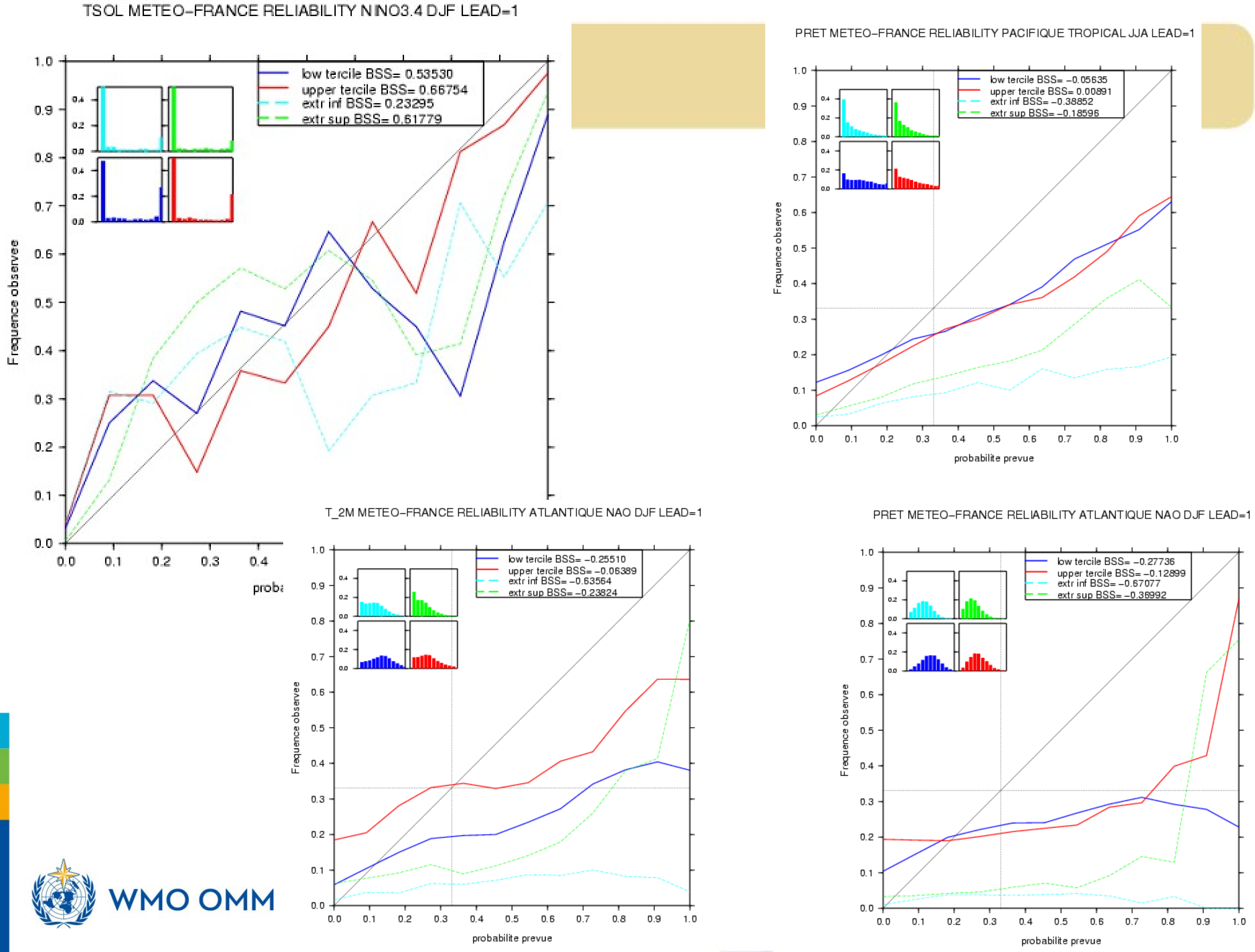
The lower the reliabity, the better it is.

Resolution

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



The greater the resolution, the better it is


TSOL METEO-FRANCE RELIABILITY NINO3.4 DJF LEAD=1
low tercile BSS= 0.53530
upper tercile BSS= 0.66754
extr inf BSS= 0.23295
extr sup BSS= 0.61779
Frequence observee
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
0.0
0.1
0.2
0.3
0.4
proba
0.0
0.2
0.4
PRET METEO-FRANCE RELIABILITY PACIFIQUE TROPICAL JJA LEAD=1
low tercile BSS= -0.05635
upper tercile BSS= 0.00891
extr inf BSS= -0.38852
extr sup BSS= -0.18596
Frequence observee
probabilite prevue
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
T_2M METEO-FRANCE RELIABILITY ATLANTIQUE NAO DJF LEAD=1
low tercile BSS= -0.25510
upper tercile BSS= -0.06389
extr inf BSS= -0.63564
extr sup BSS= -0.23824
Frequence observee
probabilite prevue
PRET METEO-FRANCE RELIABILITY ATLANTIQUE NAO DJF LEAD=1
low tercile BSS= -0.27736
upper tercile BSS= -0.12899
extr inf BSS= -0.67077
extr sup BSS= -0.36992
Frequence observee
probabilite prevue

WMO OMM