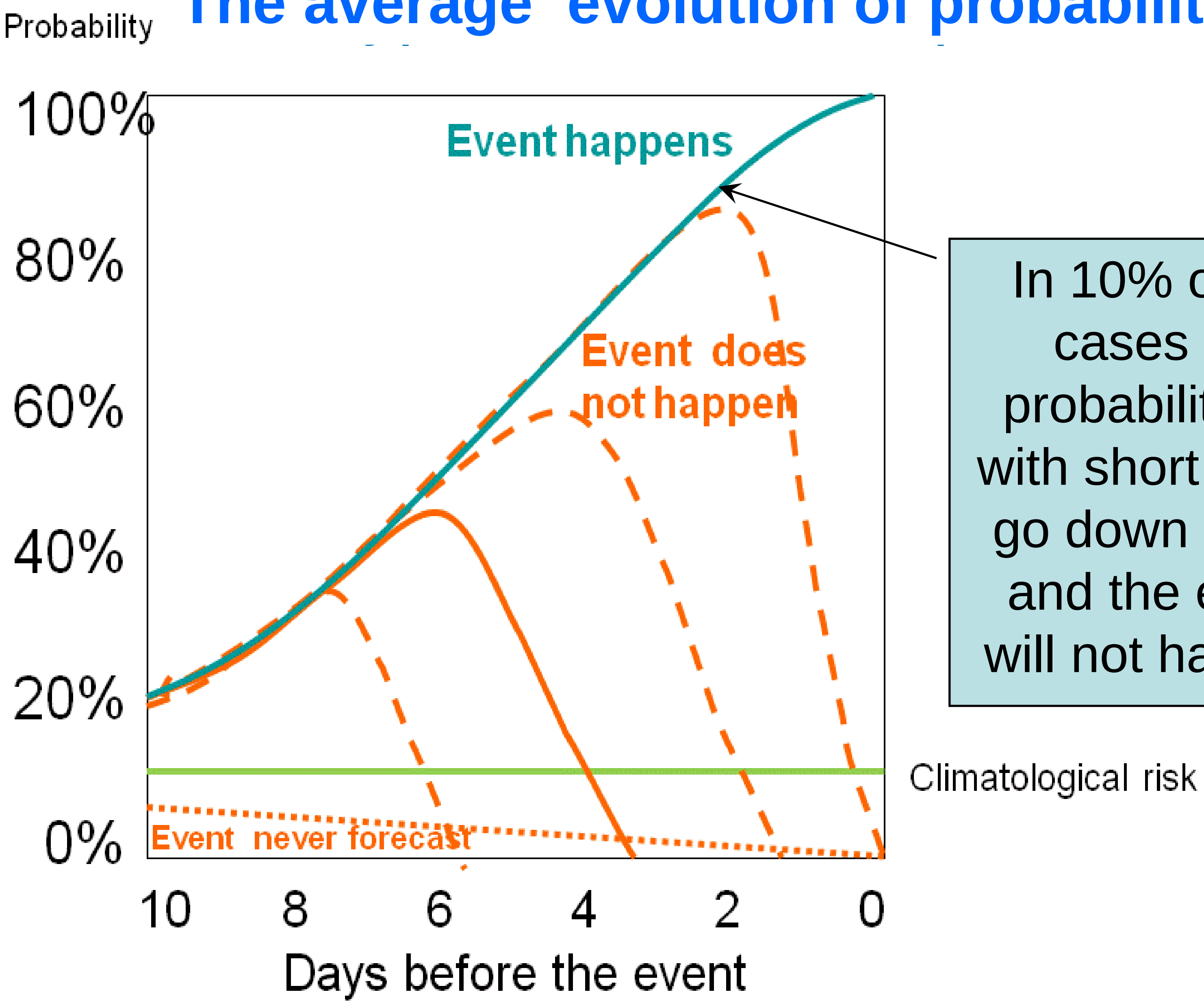
II. Frequentist probabilities

II.2 Verification of probability forecasts

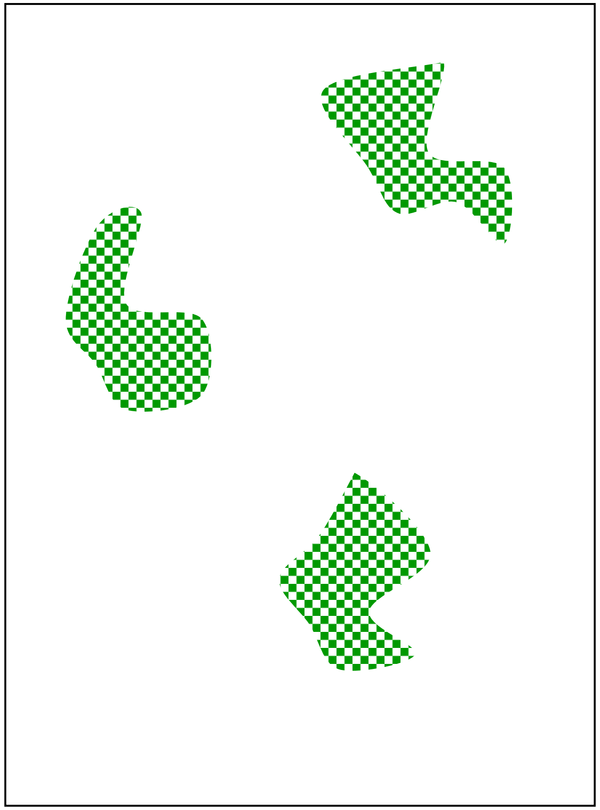
II.2.1 What is a good probability forecast?

The average evolution of probability values

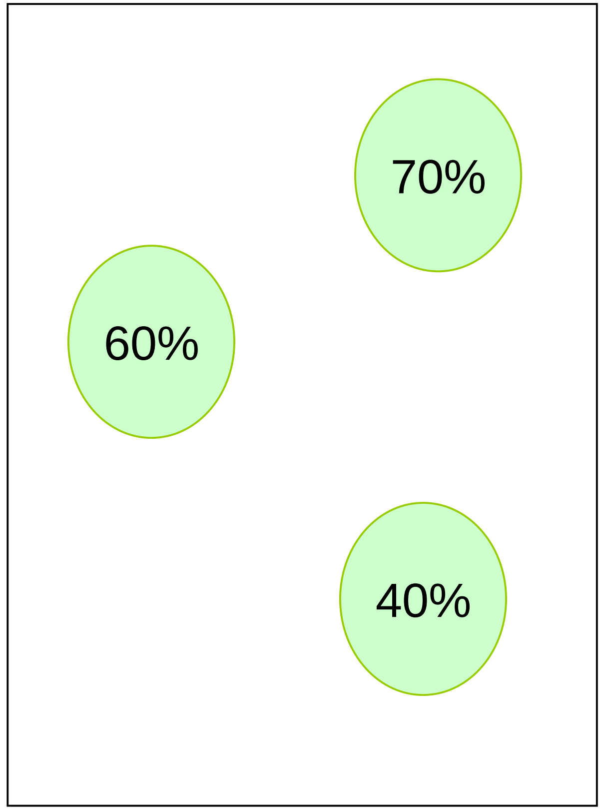


In 10% of the cases the probability will with short notice go down to 0% and the event will not happen.

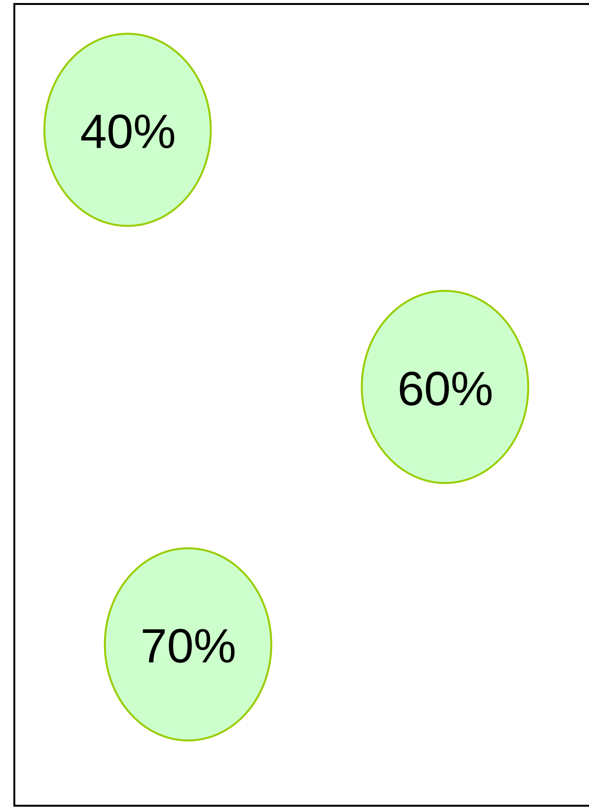
Observed rainfall (radar)



Probability forecast 1



Probability forecast 2



A scientist at a meeting showed these images:

The radar observed rain fall and the probability forecast from system 1

And then he showed the forecasts from **system 2**

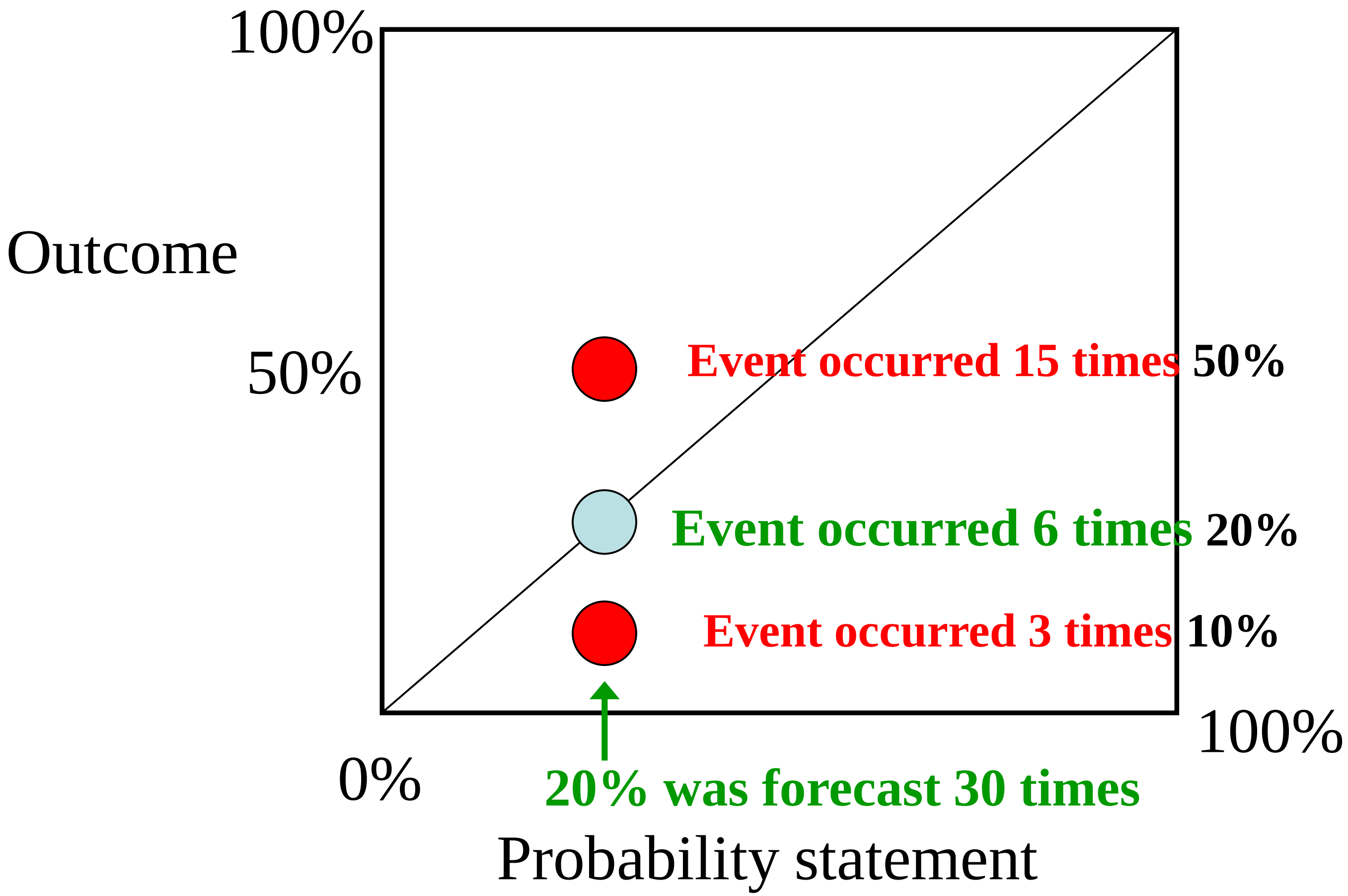
Are they worse??

Answer: -We cannot say

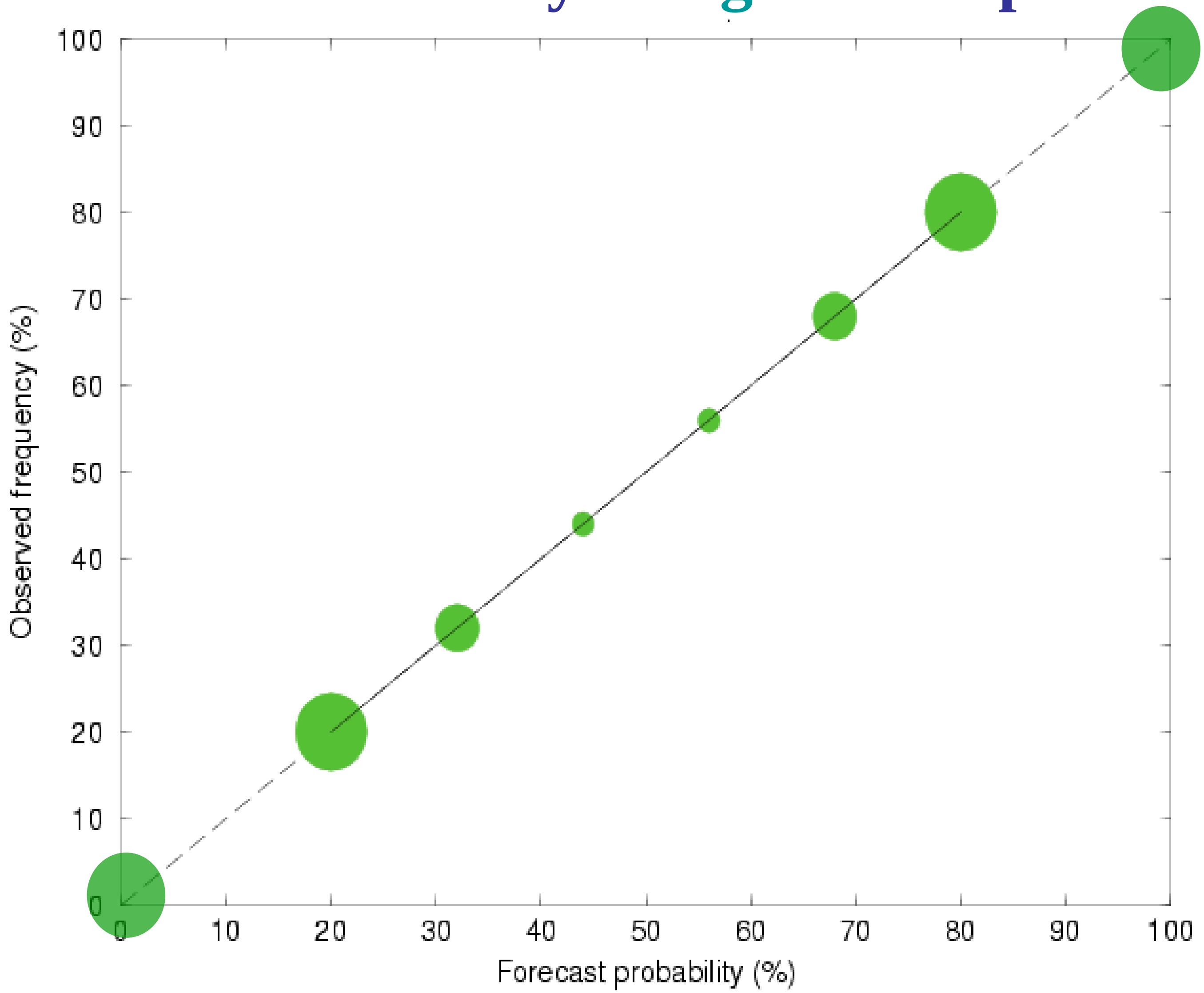
1. It is only <u>one</u> forecast

2. **If rain only fell when the probabilities were > 40% and not when they were below, something is wrong**

The reliability diagram

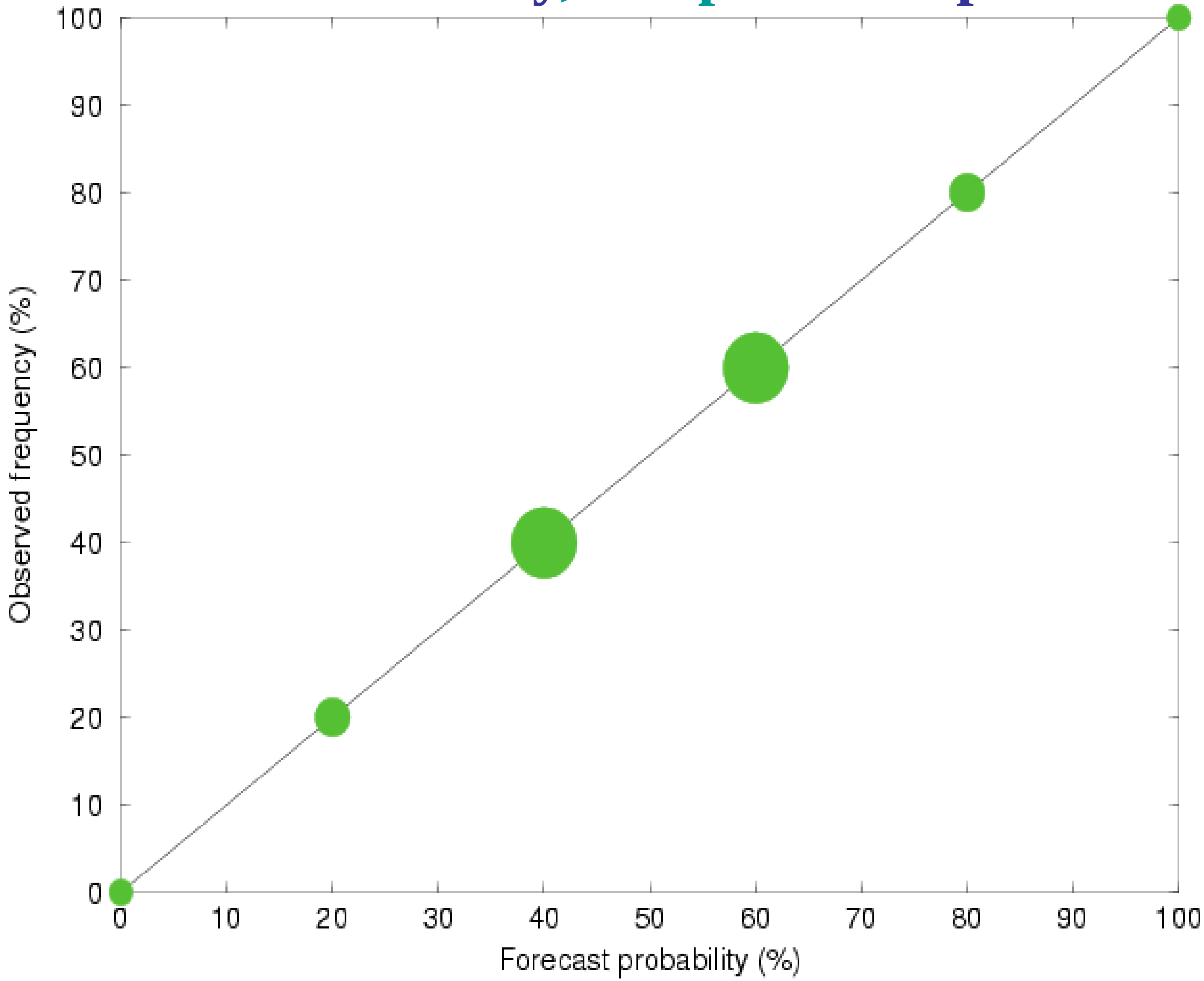


Ideal reliability and good sharpness

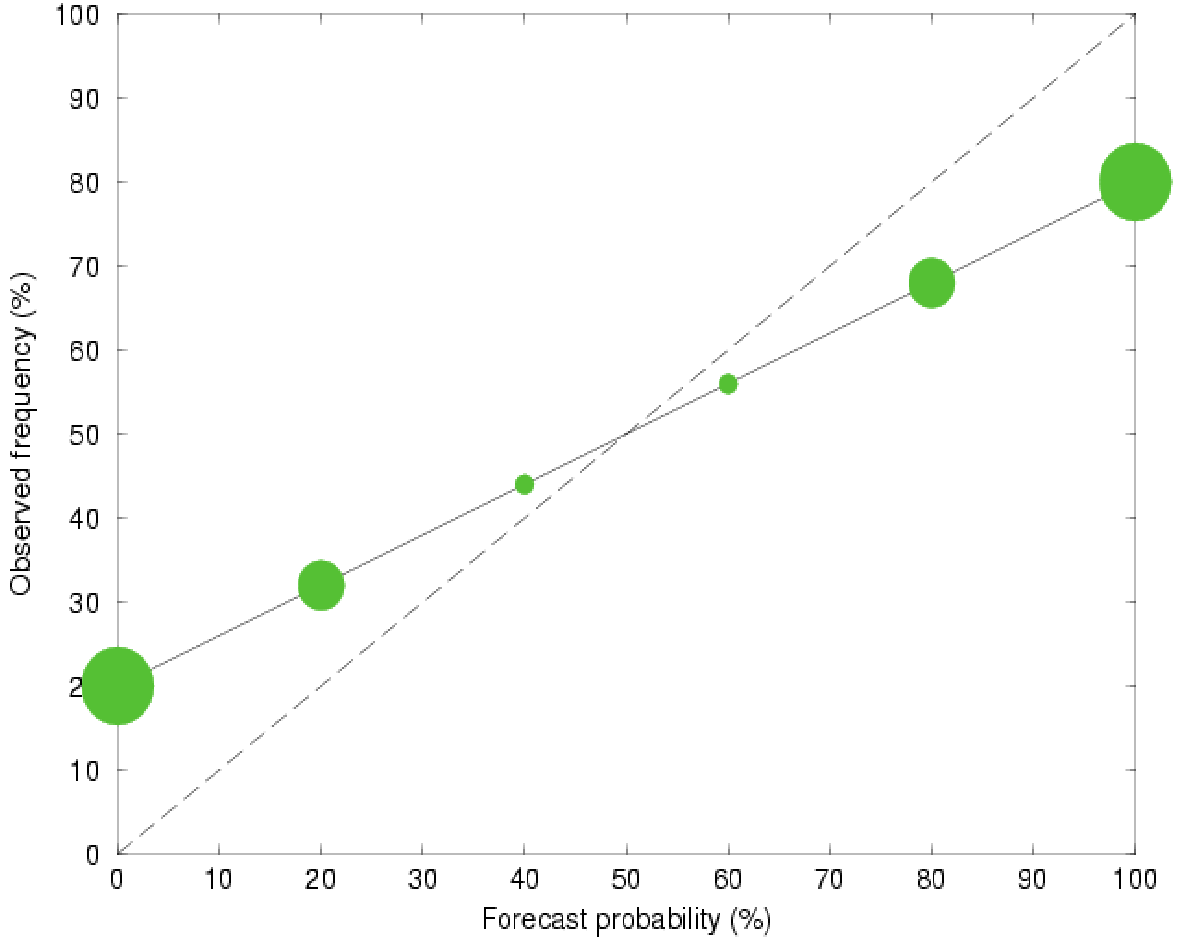


Good sharpness = forecasts draw towards 0% and 100%

Good reliability, but poor sharpness

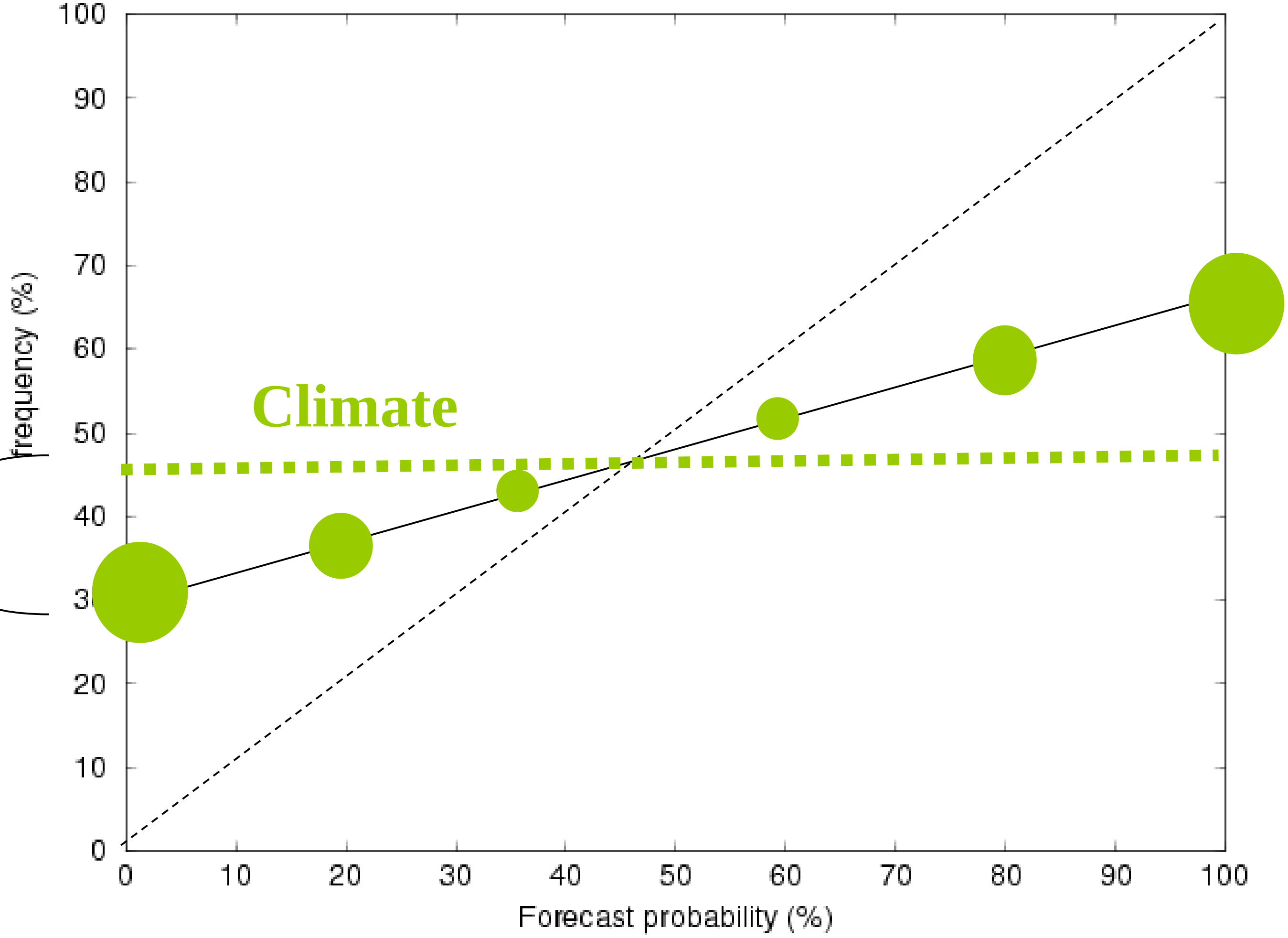


Poor reliability, but good sharpness



Do not confuse "sharpness" and "resolution"

The resolution



II.2.2 The Brier Score (BS)

The Brier score

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

Over N days

Forecast probability

Observed event (0 or 1)

BS and RMSE have identical mathematical structures

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

Over N days

Forecast probability

Observed event (0 or 1)

$$E_j^2 = \frac{1}{N}\sum_{n=1}^{N}(f_{i,j} - a_{i+j})^2$$

Over N days

Forecast (value)

Observed event (value)

The notation of the Brier score can be simplified as with RMSE

$$BS = \overline{(p - o)^2}$$

$$E^2 = \overline{(f - a)^2}$$

II.2.3 Decomposition of the Brier score

Two alternatives, Murphy (1983) which is very quoted but rarely used, or one similar to the RMSE decomposition

The Non-Murphy decomposition is identical to the RMSE one:

$$BS = \overline{(p-o)^2} =$$

$$\overline{(p-\bar{o})^2} + \overline{(o-\bar{o})^2} - \overline{2(p-\bar{o})(o-\bar{o})}$$

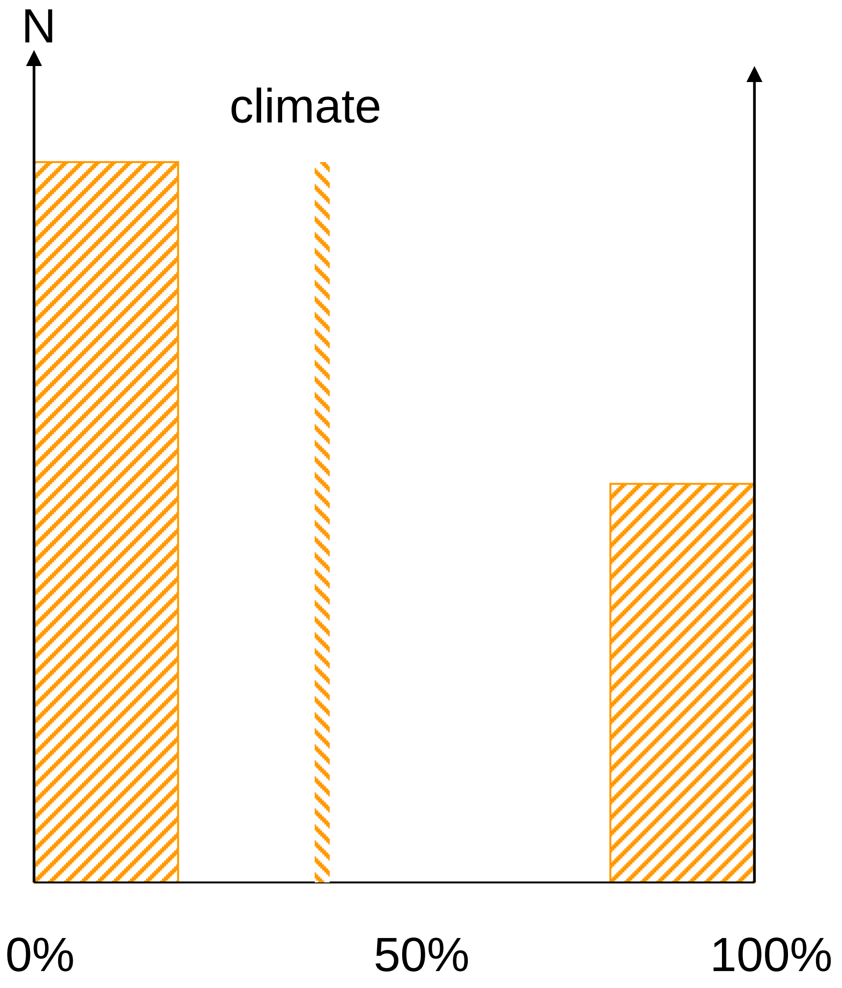
The “sharpness”

The atmospheric variability

Forecast “skill”

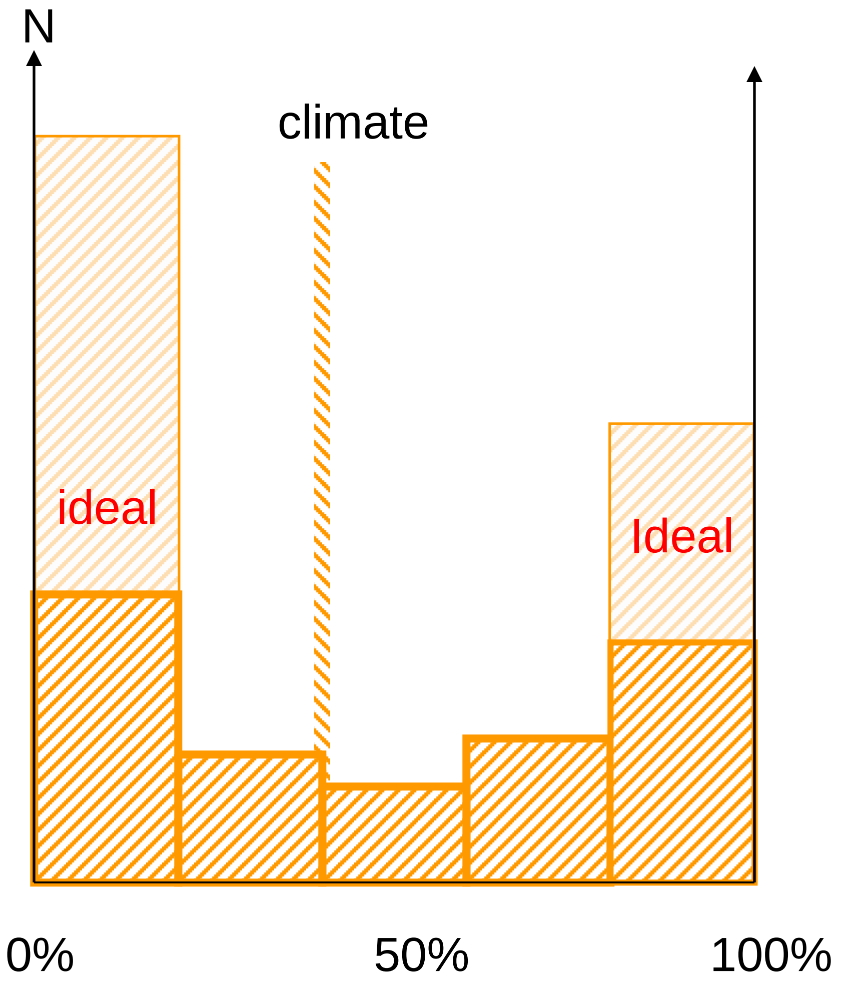
Atmospheric variability

$$\overline{(o-\overline{o})^2}$$



”Sharpness”
Model variability

$$\overline{(p-\overline{o})^2}$$



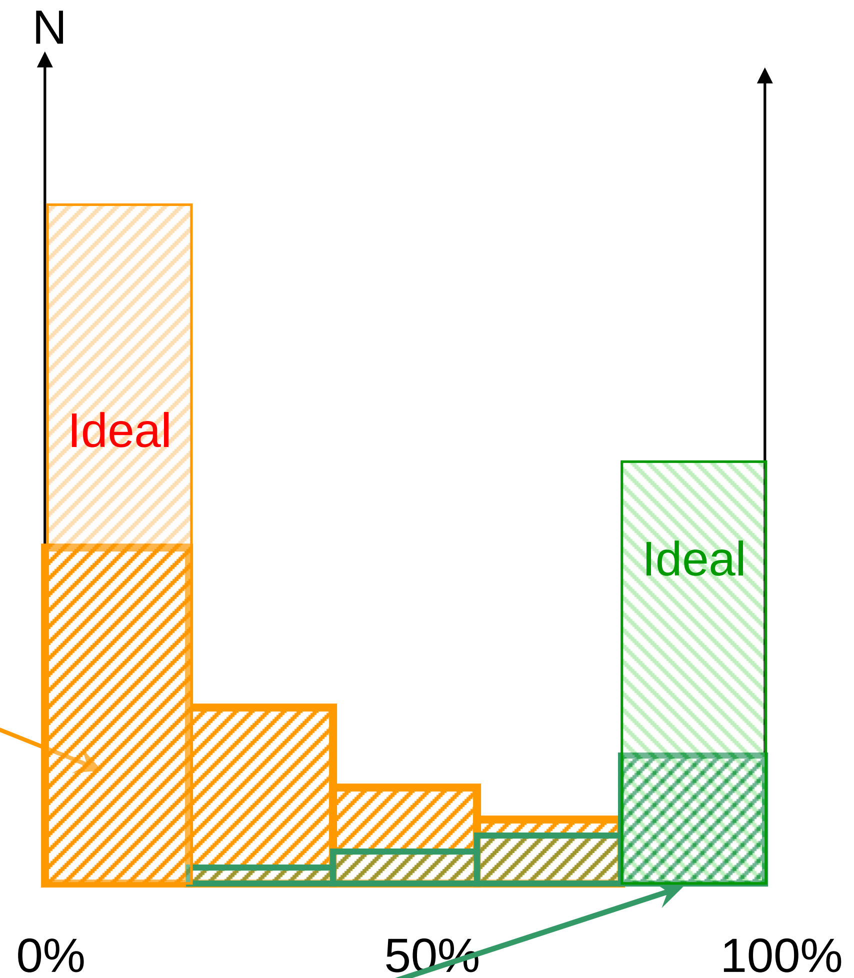
"Reliability" or "skill"

$$\overline{2(p-\bar{o})(o-\bar{o})}$$

	$0-\bar{o}$	$1-\bar{o}$
$p_1-\bar{o}$	45	5
$p_2-\bar{o}$	7	3
$p_3-\bar{o}$	5	5
$p_4-\bar{o}$	3	7
$p_5-\bar{o}$	2	18

…or graphically

$$\overline{2(p-\bar{o})(o-\bar{o})}$$



Probability forecasts **not** followed by rain

Probability forecasts followed by rain

II.2.4 Alan Murphy's decomposition

Brier Score decomposition

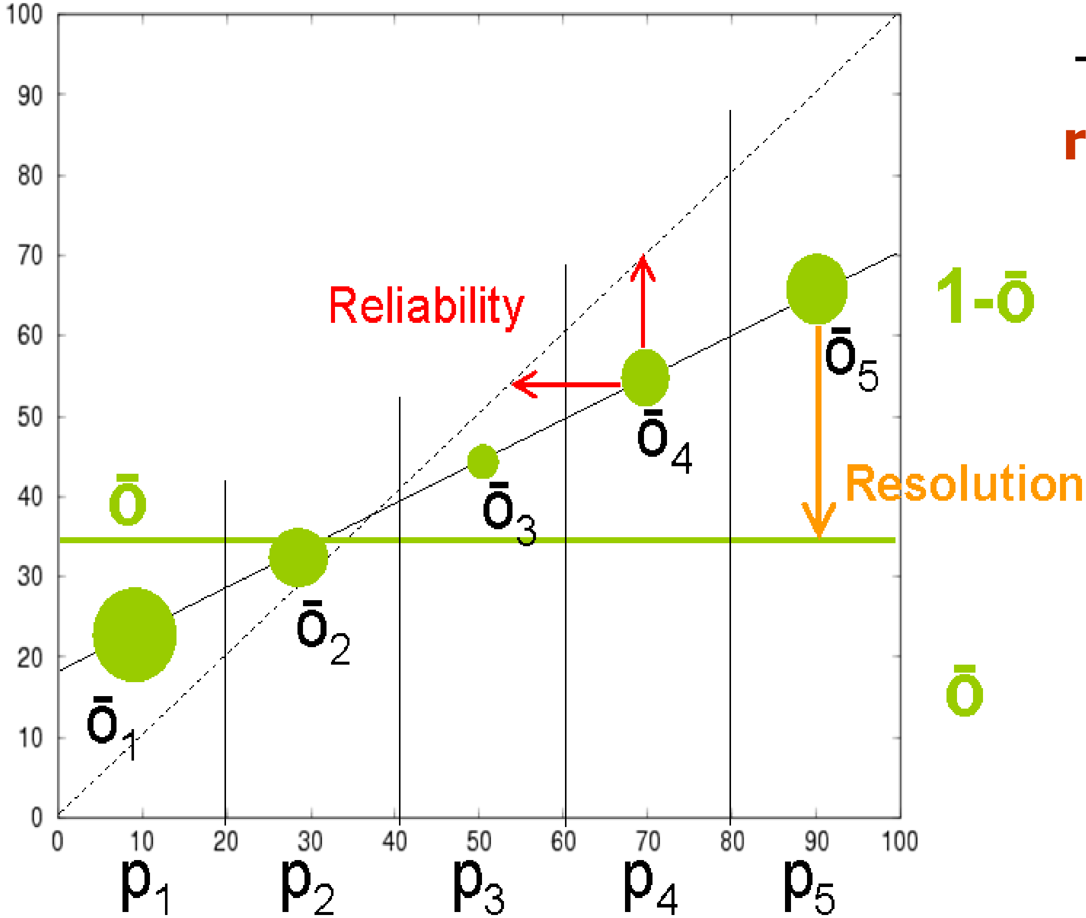
$$BS = \frac{1}{N}\sum_{k=0}^{M} N_k (f_k - \bar{o}_k)^2 - \frac{1}{N}\sum_{k=0}^{M} N_k (\bar{o}_k - \bar{o})^2 + \bar{o}(1-\bar{o})$$

	$0-\bar{o}$	$1-\bar{o}$
$p_1-\bar{o}$	45	5
$p_2-\bar{o}$	3	7
$p_3-\bar{o}$	5	5
$p_4-\bar{o}$	7	3
$p_5-\bar{o}$	2	18

The first term is a **reliability** measure:

For perfectly reliable forecasts, the sub-sample relative frequency is exactly equal to the forecast probability in each sub-sample.

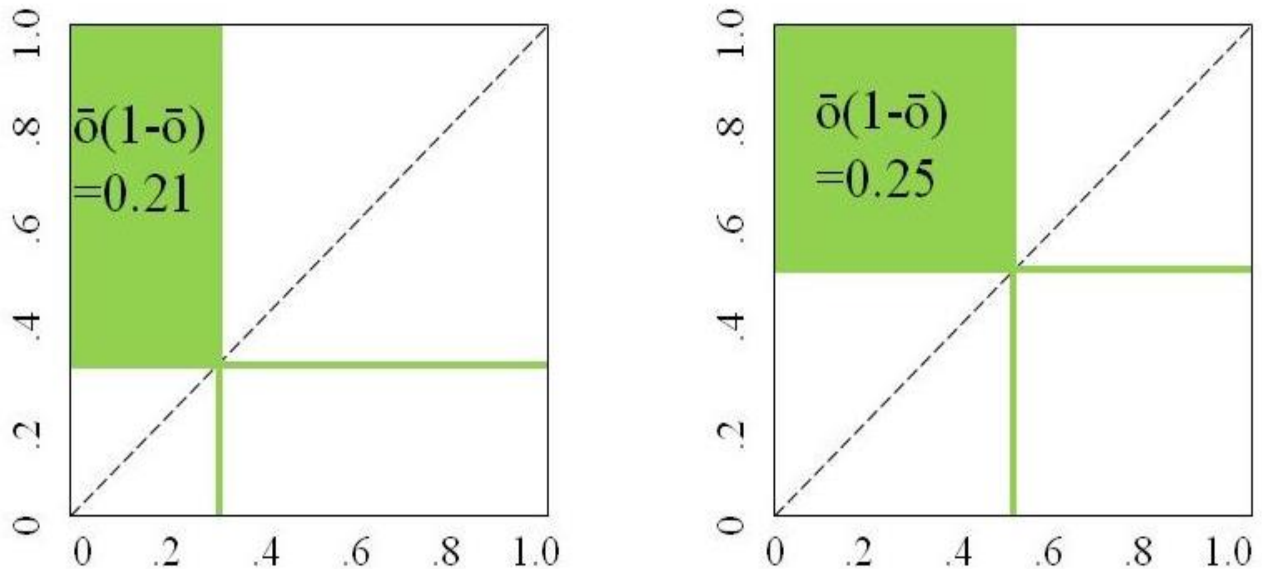
$$BS = \frac{1}{N}\sum_{k=0}^{M} N_k (f_k - \bar{o}_k)^2 - \frac{1}{N}\sum_{k=0}^{M} N_k (\bar{o}_k - \bar{o})^2 + \bar{o}(1-\bar{o})$$



The second term is a **resolution** measure:

$$BS = \frac{1}{N}\sum_{k=0}^{M} N_k (f_k - \overline{o}_k)^2 - \frac{1}{N}\sum_{k=0}^{M} N_k (\overline{o}_k - \overline{o})^2 + \overline{o}(1-\overline{o})$$

The **uncertainty term** ranges from 0 to 0.25. If the event either always occurs or never occurs, then there is high certainty. With a 50-50 probability it is most uncertain



ō = 0.3 yields less “uncertainty” (0.21) than ō = 0.5 (0.25)

Compare with a sack of balls with two colours with proportions ō and 1-ō where “certainty” = $ō^2 + (1-ō)^2$

II.2.5 Pitfalls with the Brier Score

The BS will appear to improve if the sharpness gets worse.

A contest between a real and fake doctor trying to forecast the sex of not yet born children.

The fake doctor will score BS = 0.5 just by guessing.

BS = 1 in 50% and
BS = 0 in 50%

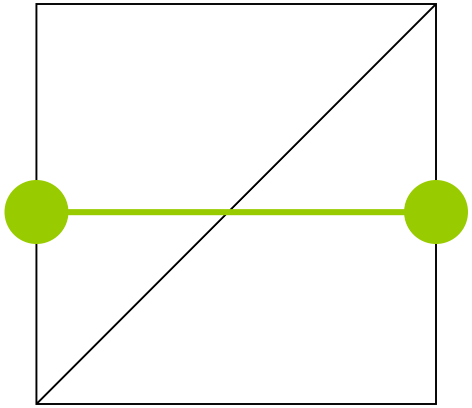
If the real doctor is 65% correct in his forecasts he will score BS = 0.35.

BS = 1 in 35% and
BS = 0 in 65%

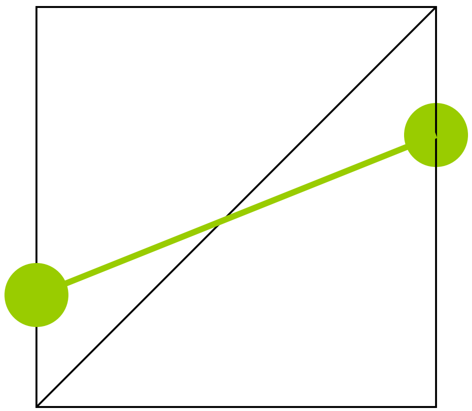
By saying "fifty-fifty" in 60% of the cases the fake doctor can "improve" his score to exactly the same BS = 0.35.

BS = 1 in 20%
BS = 0 in 20%
BS = 0.25 in 60%

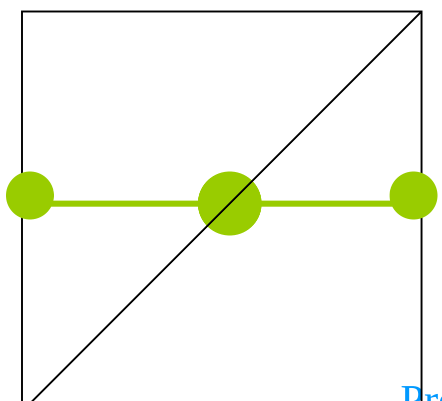
0.2 + 0.15 = 0.35



The guessing hoaxer's reliability diagram. Brier score = **0.50**



The rather skilful (60% hit rate) scientist's reliability diagram. Brier score = **0.35**



The hoaxer's "improved" reliability diagram. He is still guessing but the Brier score has decreased to **0.35** since he increased a "useless" reliability

I.2.6 Extensions of the Brier Score

Brier skill score

A new ”Brier Skill Score” with climate as reference would be

$$\boxed{BSS_{new} = \frac{\overline{(p-\bar{o})(o-\bar{o})}}{\overline{(o-\bar{o})^2}}}$$

In analogy with the Anomaly Correlation Coefficient

$$ACC = \frac{\overline{(f-c)(a-c)}}{\overline{(a-c)^2}}$$

Instead the following definition has been agreed

Brier skill score

$$BSS = (BS_{ref} – BS)/BS_{ref}$$

Rank probability score (RPS)

It is just the Brier score

$$BS = \overline{(p - o)^2}$$

applied for different thresholds, defining new probabilities, and then integrating or summing up

END