



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)



Ensemble Prediction System

- 1 control run + 50 perturbed runs (T_L639 L62)
 - added dimension of ensemble members
 - $f(x,y,z,t,e)$
- How do we deal with added dimension when
 - interpreting, verifying and diagnosing EPS output?

Transition from deterministic (yes/no) to probabilistic



Assessing the quality of a forecast

- The forecast indicated 10% probability for rain
- It did rain on the day
- Was it a good forecast?
 - Yes
 - No
 - I don't know



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,...)?

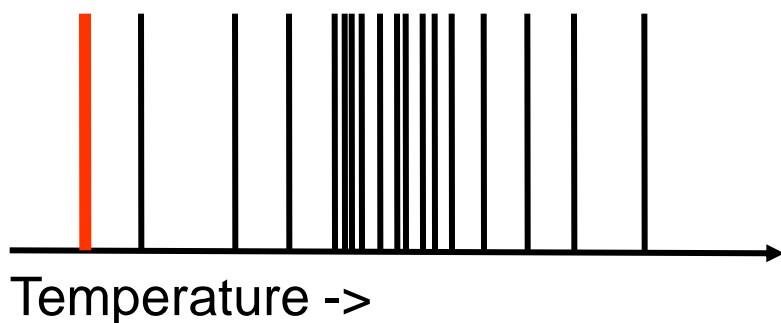
* Note that terms like consistency, reliability etc. are not always well defined in verification theory and can be used with different meanings in other contexts



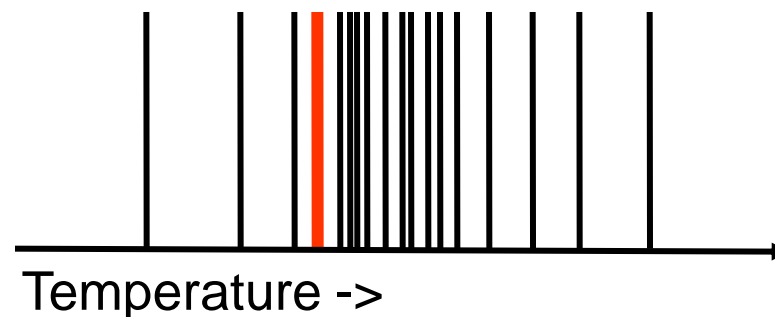
Rank Histogram

- Rank Histograms assess 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 1 case



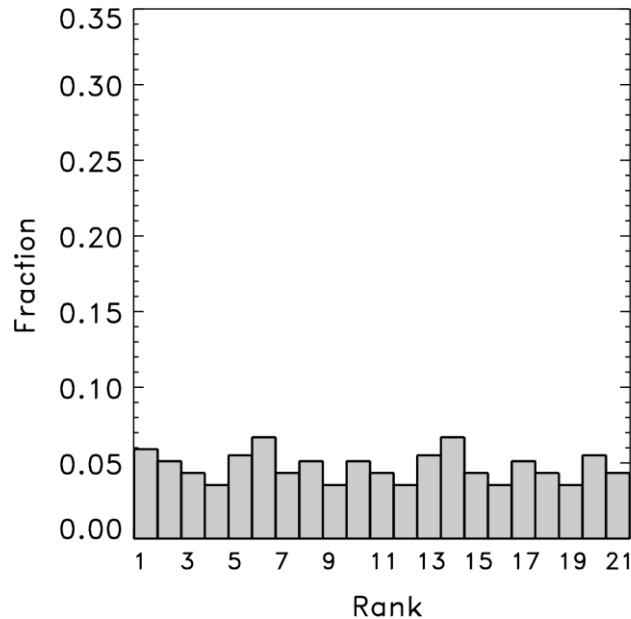
Rank 4 case





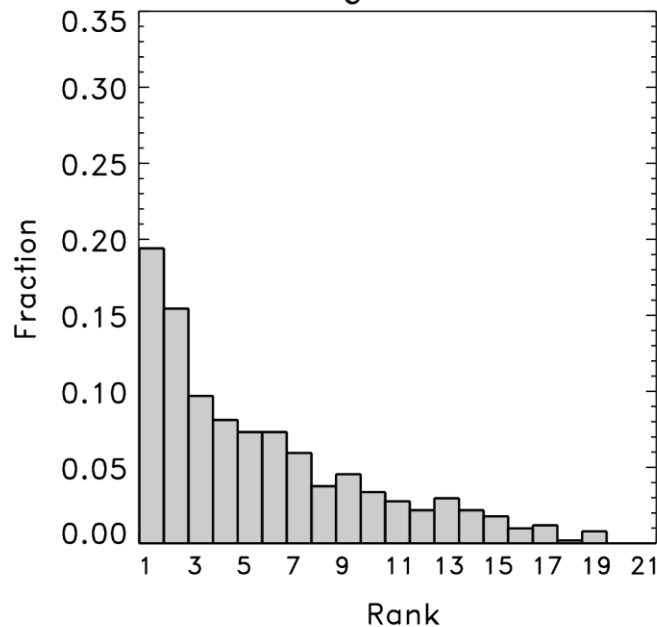
Rank Histograms

OK



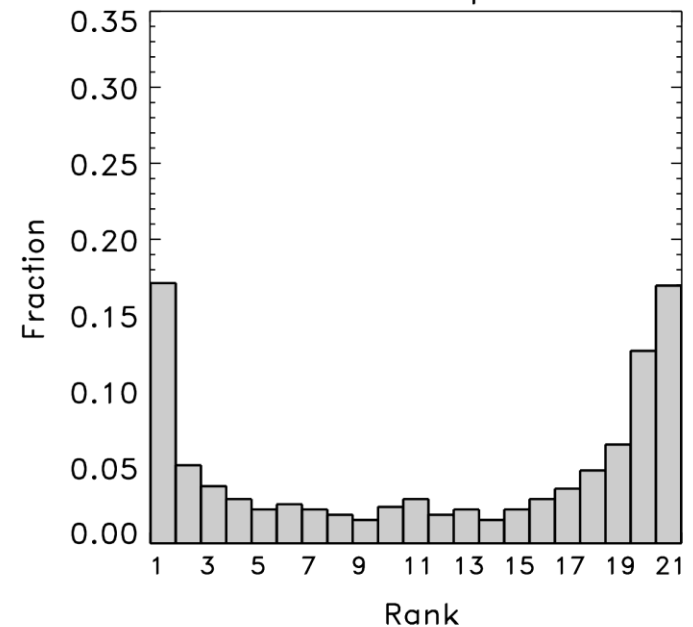
OBS is indistinguishable from any other ensemble member

High Bias



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

Too Little Spread



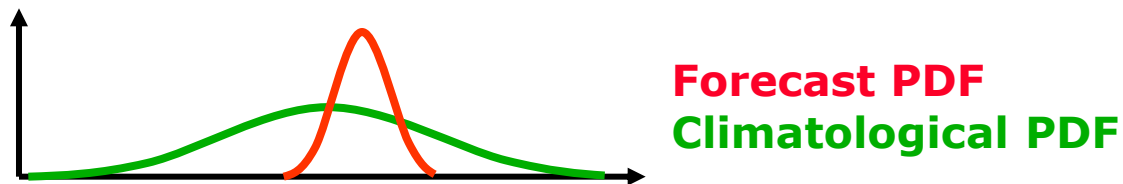
OBS is too often outside the ensemble spread

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

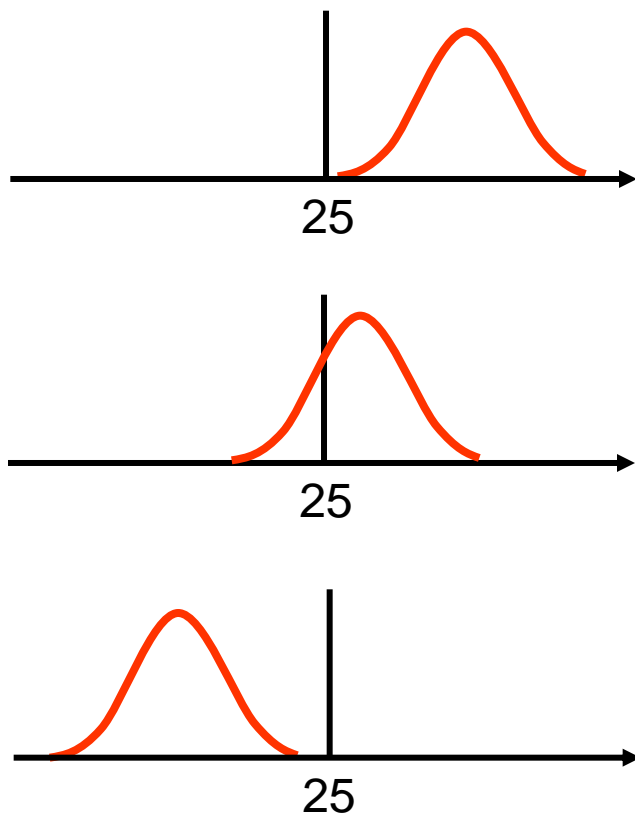




Reliability Diagram

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?



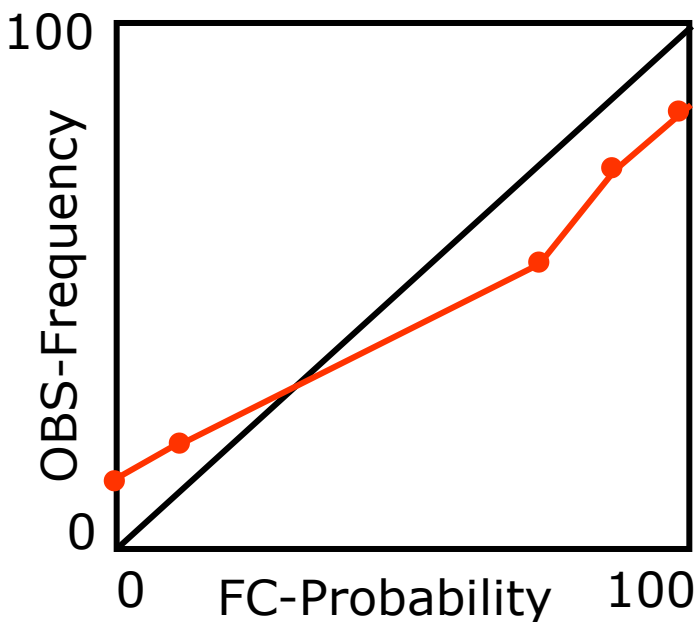
FC Prob.	# FC	OBS-Frequency (perfect model)	OBS-Frequency (imperfect model)
100%	8000	8000 (100%)	7200 (90%)
90%	5000	4500 (90%)	4000 (80%)
80%	4500	3600 (80%)	3000 (66%)
....
....
....
10%	5500	550 (10%)	800 (15%)
0%	7000	0 (0%)	700 (10%)



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?

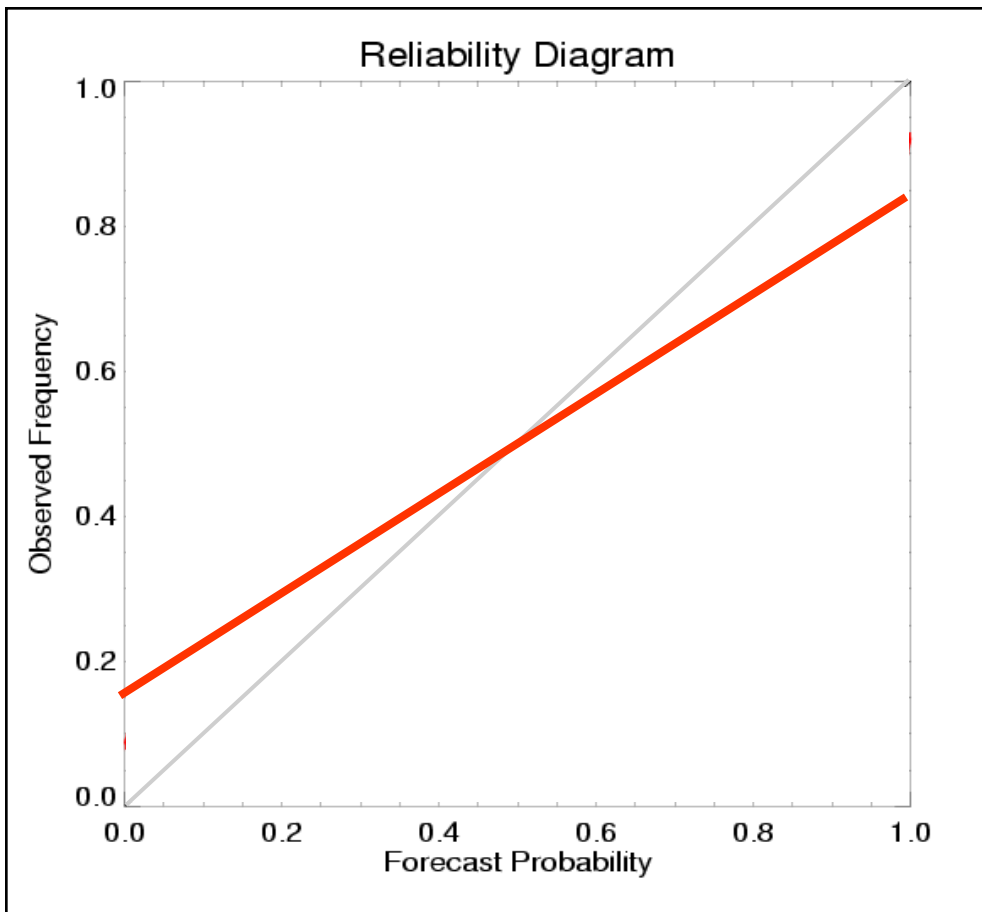


FC Prob.	# FC	OBS-Frequency (perfect model)	OBS-Frequency (imperfect model)
100%	8000	8000 (100%)	7200 (90%)
90%	5000	4500 (90%)	4000 (80%)
80%	4500	3600 (80%)	3000 (66%)
....
....
....
10%	5500	550 (10%)	800 (15%)
0%	7000	0 (0%)	700 (10%)

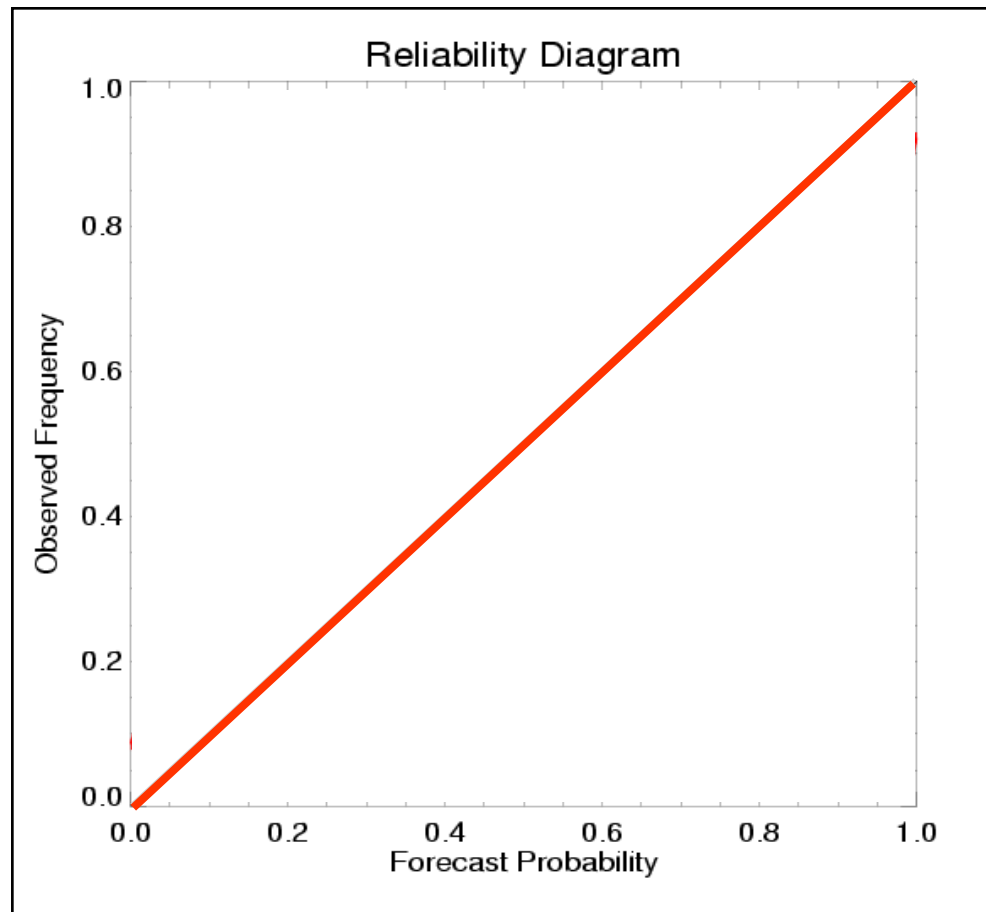


Reliability Diagram

over-confident model



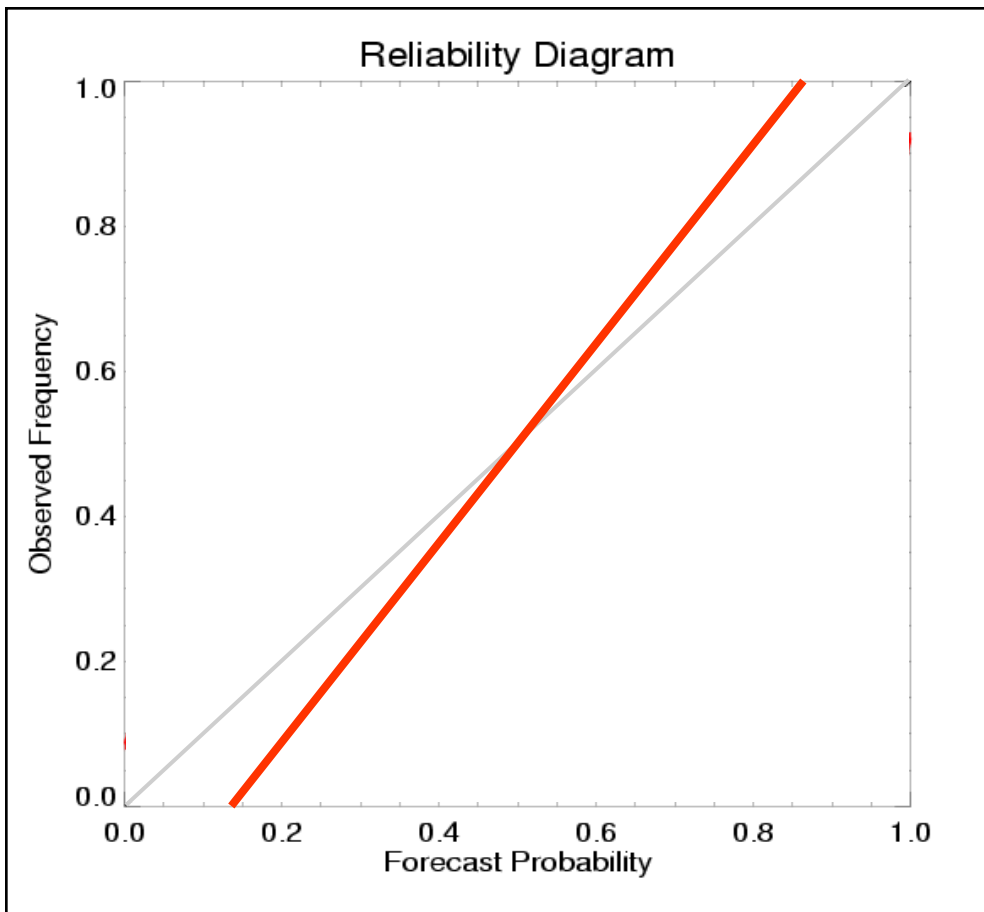
perfect model



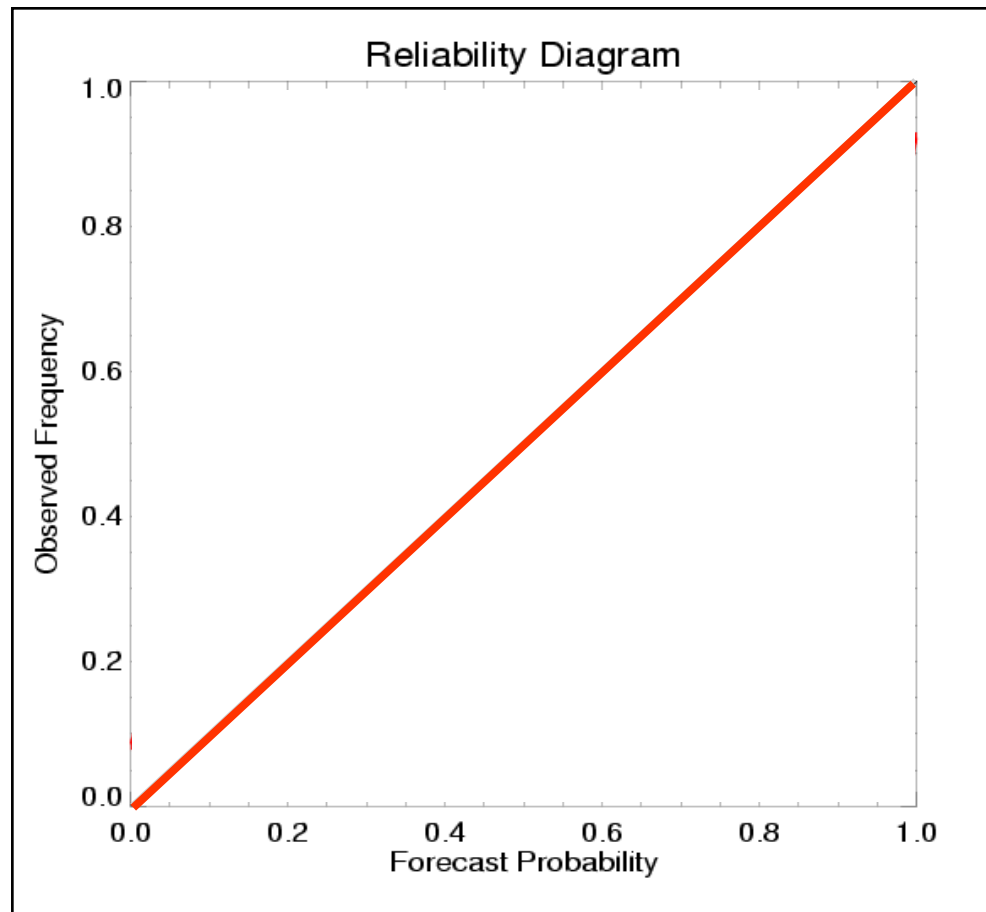


Reliability Diagram

under-confident model



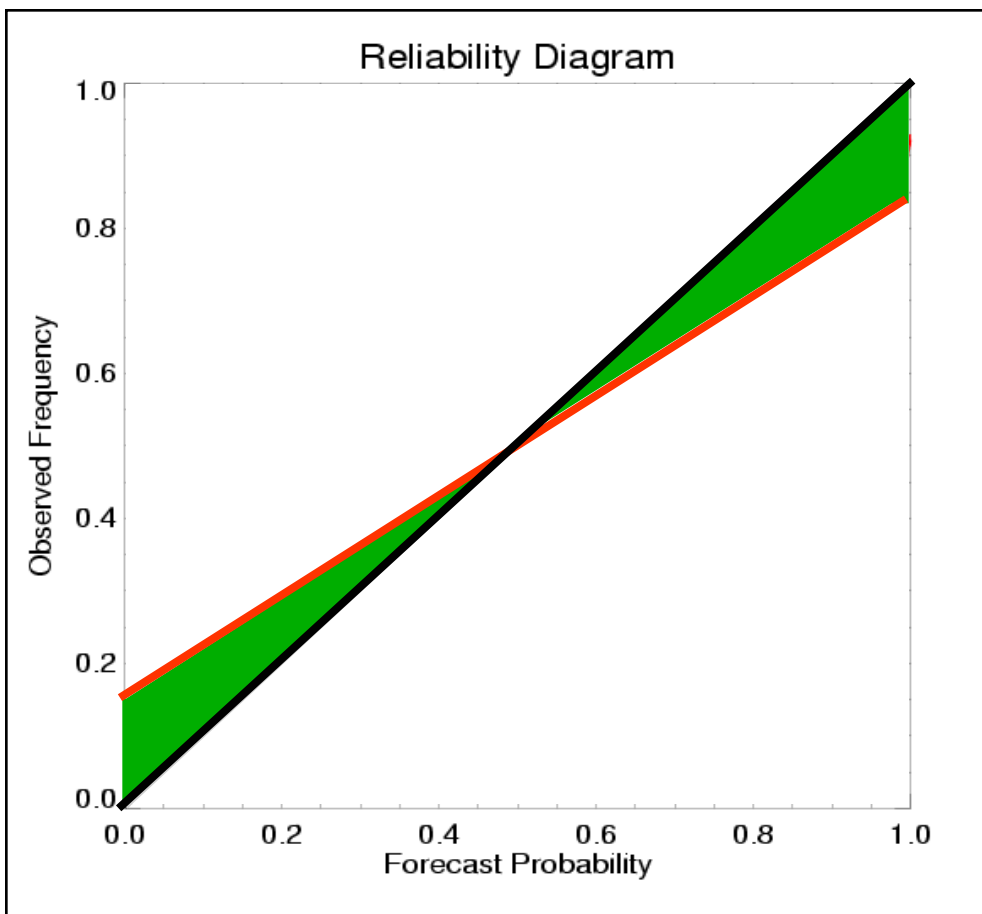
perfect model



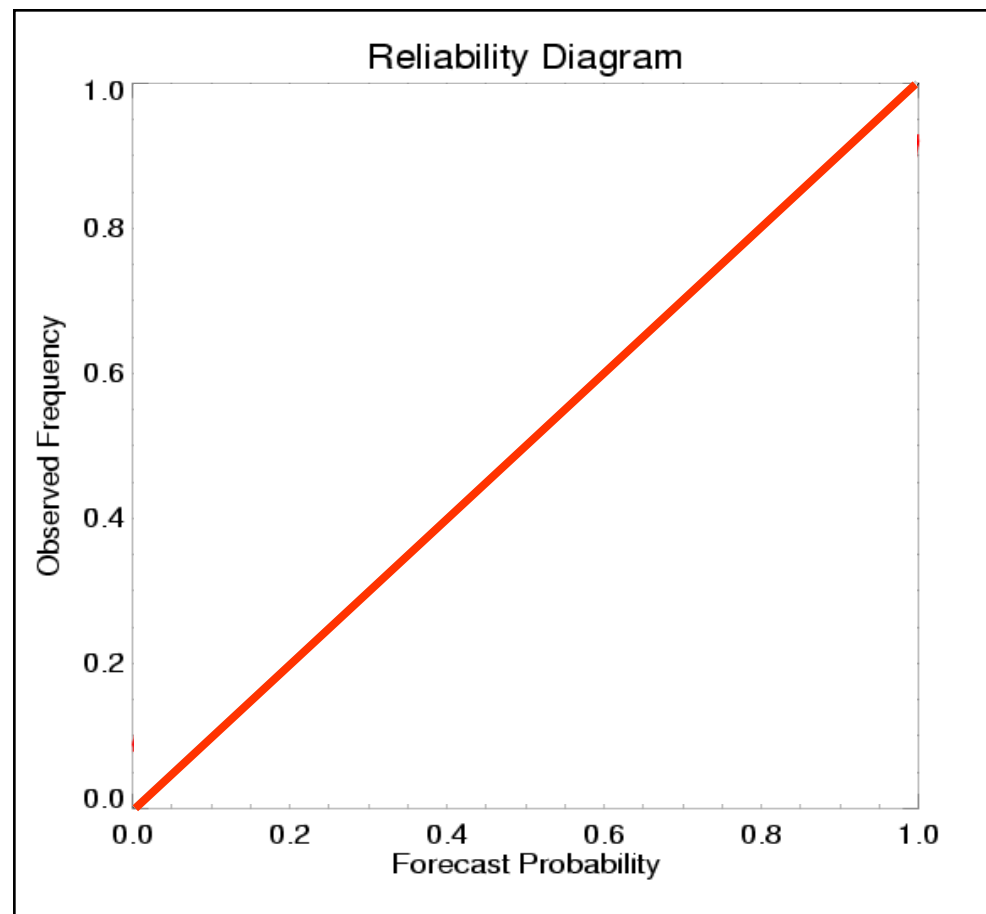


Reliability diagram

 Reliability score (the smaller, the better)



imperfect model



perfect model



Components of the Brier Score

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

N = total number of cases

I = number of probability bins

n_i = number of cases in probability bin i

p_i = forecast probability in probability bin i

o_i = frequency of event being observed when forecasted with p_i

- Reliability: forecast probability vs. observed relative frequencies



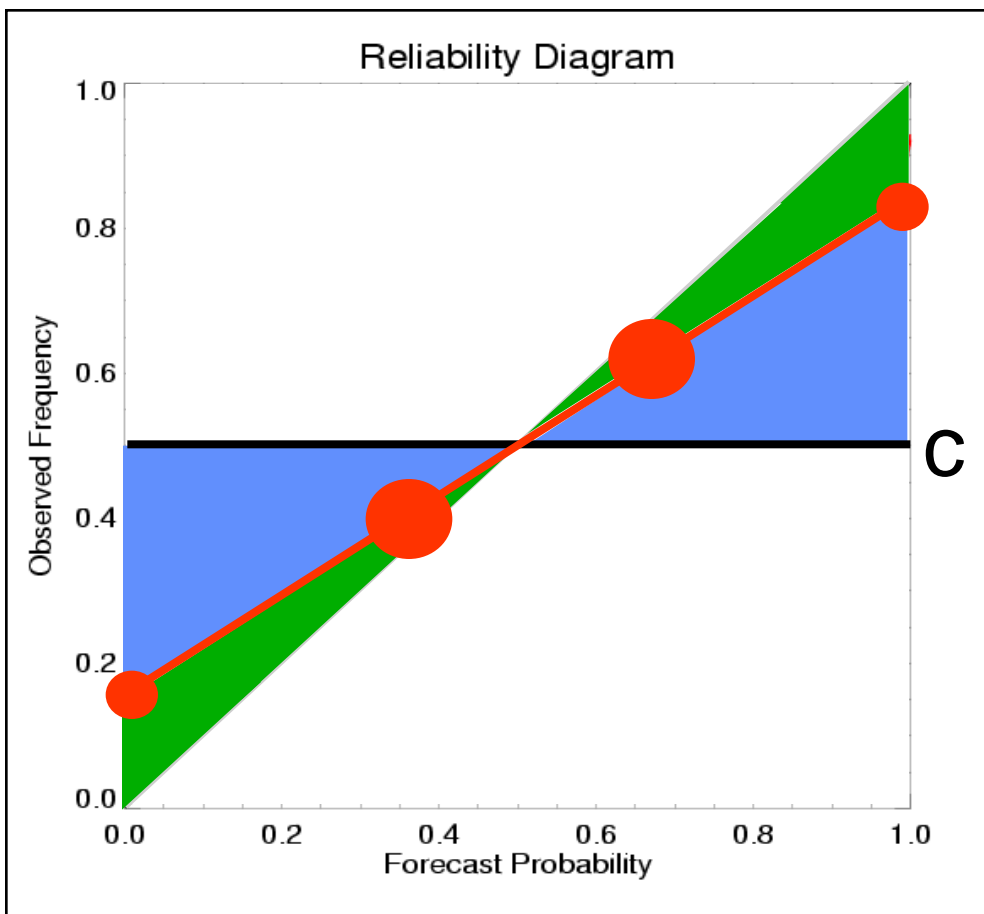
Reliability diagram



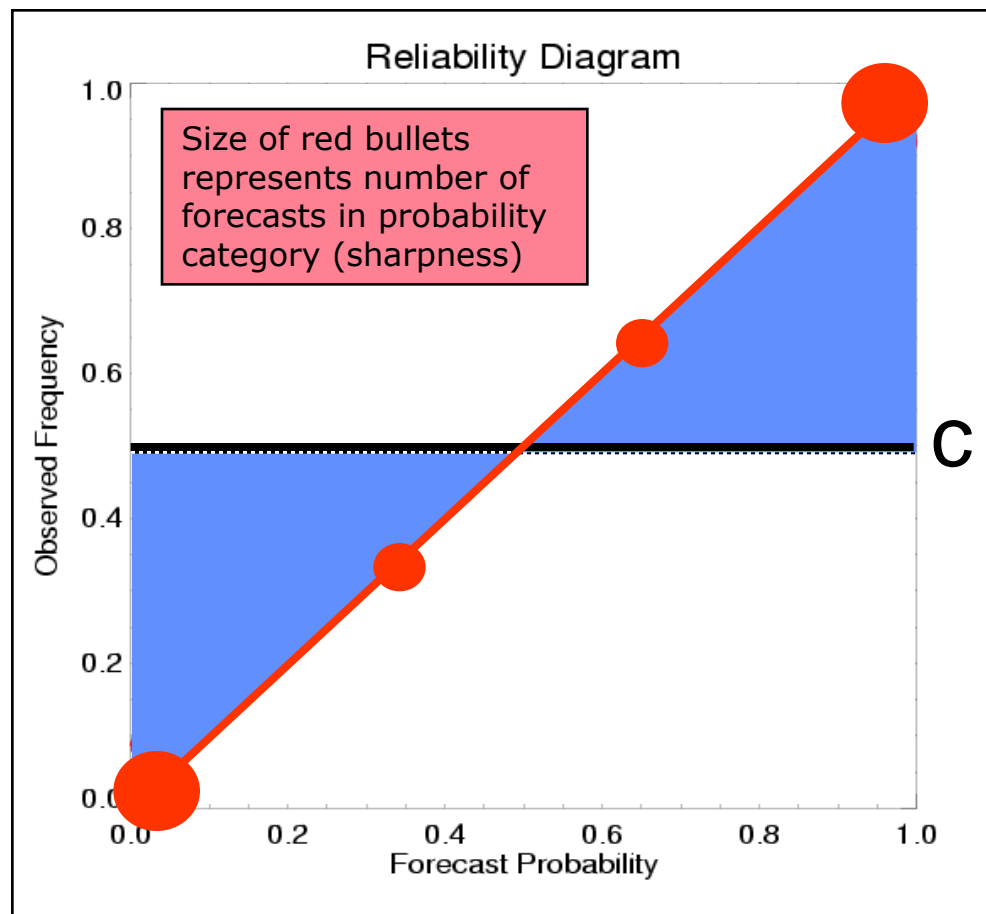
Reliability score (the smaller, the better)



Resolution score (the bigger, the better)



Poor resolution



Good resolution



Components of the Brier Score

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

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

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

N = total number of cases

I = number of probability bins

n_i = number of cases in probability bin i

p_i = forecast probability in probability bin i

o_i = frequency of event being observed when forecasted with p_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

$$\text{Brier Score} = \text{Reliability} - \text{Resolution} + \text{Uncertainty}$$



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

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

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

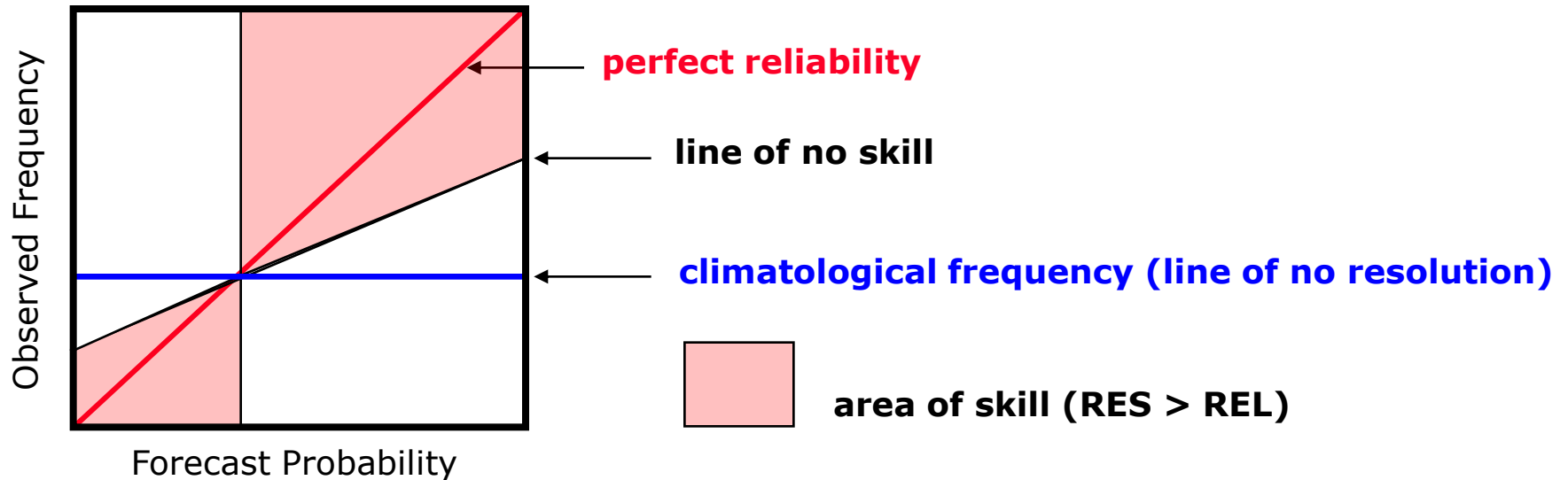
- **positive** (**negative**) BSS ➤ **better** (**worse**) than reference



Brier Skill Score & Reliability Diagram

- How to construct the area of positive skill?

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





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 event, and the systems gets it right?

Brier

Skill Score

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

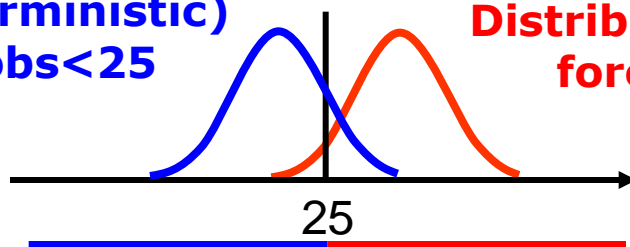


Discrimination

- Until now, we looked at the question:
 - If the forecast system predicts x , what is the observation y ?
- When we are interested in the ability of a forecast system to discriminate between events and non-events, we investigate the question:
 - If the event y occurred, what was the forecast x ?
- Based on signal-detection theory, the Relative Operating Characteristic (ROC) measures this discrimination ability

**Distribution of (deterministic)
forecasts when obs < 25**

**Distribution of (deterministic)
forecasts when obs ≥ 25**

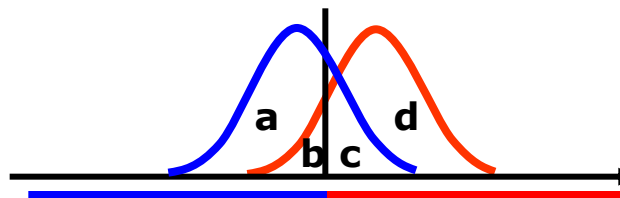




Verification of two category (yes/no) situations

- Compute 2 x 2 contingency table:
(for a set of cases)

		Event observed		total
		Yes	No	
Event forecasted	Yes	a	b	a+b
	No	c	d	c+d
total		a+c	b+d	a+b+c+d=n



- Event Probability: $s = (a+c) / n$
- Probability of a Forecast of occurrence: $r = (a+b) / n$
- Frequency Bias: $B = (a+b) / (a+c)$
- Proportion Correct: $PC = (a+d) / n$



Example of Finley Tornado Forecasts (1884)

- Compute 2 x 2 contingency table:
(for a set of cases)

		Event observed		total
		Yes	No	
Event forecasted	Yes	28	72	100
	No	23	2680	2703
total		51	2752	2803

- Event Probability: $s = (a+c) / n = 51/2803 = 0.018$
- Probability of a Forecast of occurrence: $r = (a+b) / n = 100/2803 = 0.036$
- Frequency Bias: $B = (a+b) / (a+c) = 100/51 = 1.961$
- Proportion Correct: $PC = (a+d) / n = 2708/2803 = 0.966$

96.6% Accuracy



Example of Finley Tornado Forecasts (1884)

- Compute 2 x 2 contingency table:
(for a set of cases)

		Event observed		total
		Yes	No	
Event forecasted	Yes	0	0	0
	No	51	2752	2803
total		51	2752	2803

- Event Probability: $s = (a+c) / n = 51/2803 = 0.018$
- Probability of a Forecast of occurrence: $r = (a+b) / n = 0/2803 = 0.0$
- Frequency Bias: $B = (a+b) / (a+c) = 0/51 = 0.0$
- Proportion Correct: $PC = (a+d) / n = 2752/2803 = 0.982$

98.2% Accuracy!



Some Scores and Skill Scores

Score	Formula	Finley (original)	Finley (never fc T.)	Finley (always fc. T.)
Proportion Correct	$PC=(a+d)/n$	0.966	0.982	0.018
Threat Score	$TS=a/(a+b+c)$	0.228	0.000	0.018
Odds Ratio	$\Theta=(ad)/(bc)$	45.3	-	-
Odds Ratio Skill Score	$Q=(ad-bc)/(ad+bc)$	0.957	-	-
Heidke Skill Score	$HSS=2(ad-bc)/((a+c)(c+d)+(a+b)(b+d))$	0.355	0.0	0.0
Peirce Skill Score	$PSS=(ad-bc)/(a+c)(b+d)$	0.523	0.0	0.0
Clayton Skill Score	$CSS=(ad-bc)/(a+b)(c+d)$	0.271	-	-
Gilbert Skill Score (ETS)	$GSS=(a-a_{ref})/(a-a_{ref}+b+c)$ $a_{ref} = (a+b)(a+c)/n$	0.216	0.0	0.0



Verification of two category (yes/no) situations

- Compute 2 x 2 contingency table:
(for a set of cases)

		Event observed		total
		Yes	No	
Event forecasted	Yes	a	b	a+b
	No	c	d	c+d
total		a+c	b+d	a+b+c+d=n

- Event Probability: $s = (a+c) / n$
- Probability of a Forecast of occurrence: $r = (a+b) / n$
- Frequency Bias: $B = (a+b) / (a+c)$
- Hit Rate: $H = a / (a+c)$
- False Alarm Rate: $F = b / (b+d)$
- False Alarm Ratio: $FAR = b / (a+b)$



Example of Finley Tornado Forecasts (1884)

- Compute 2 x 2 contingency table:
(for a set of cases)

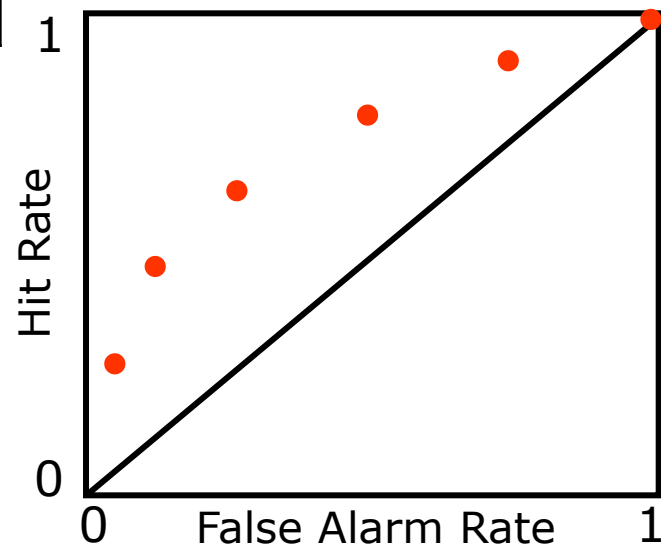
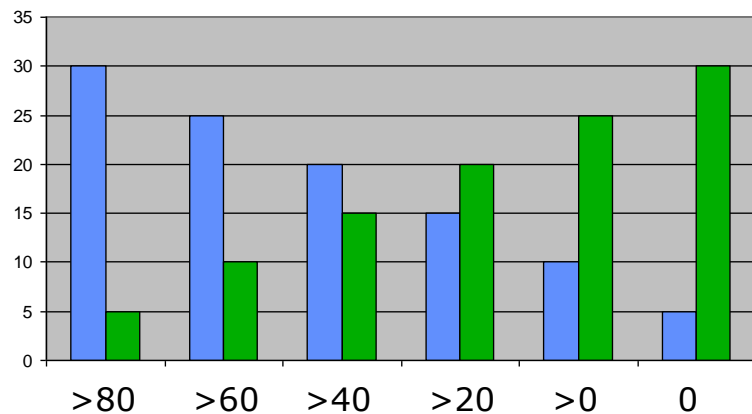
		Event observed		total
		Yes	No	
Event forecasted	Yes	28	72	100
	No	23	2680	2703
total		51	2752	2803

- Event Probability: $s = (a+c) / n = 0.018$
- Probability of a Forecast of occurrence: $r = (a+b) / n = 0.036$
- Frequency Bias: $B = (a+b) / (a+c) = 1.961$
- Hit Rate: $H = a / (a+c) = 0.549$
- False Alarm Rate: $F = b / (b+d) = 0.026$
- False Alarm Ratio: $FAR = b / (a+b) = 0.720$



Extension of 2 x 2 contingency table for prob. FC

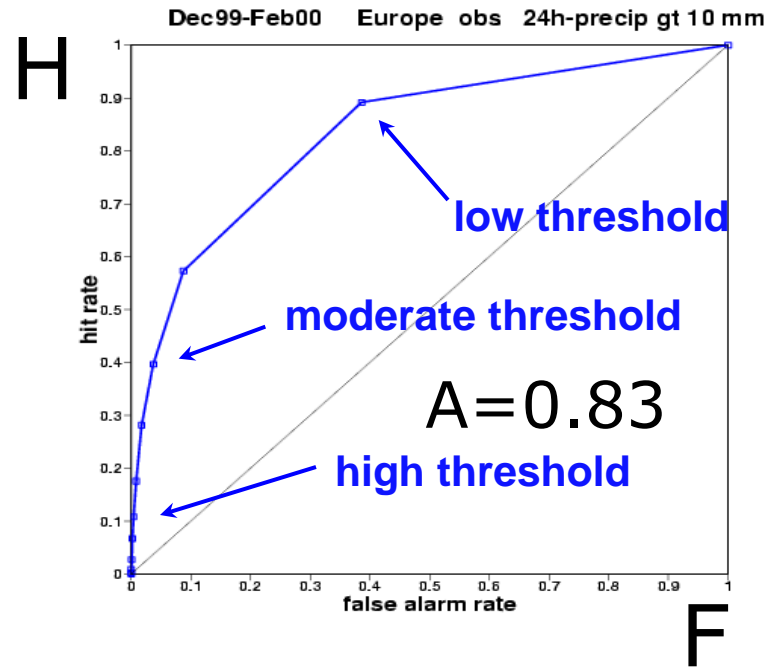
		Event observed	
		Yes	No
Event forecasted	>80% - 100%	30	5
	>60% - 80%	25	10
	>40% - 60%	20	15
	>20% - 40%	15	20
	>0% - 20%	10	25
	0%	5	30
	total	105	105





ROC curve

- ROC curve is plot of H against F for range of probability thresholds



- ROC area (area under the ROC curve) is skill measure
 $A=0.5$ (no skill), $A=1$ (perfect deterministic forecast)
- ROC curve is independent of forecast bias, i.e. represents potential skill
- ROC is conditioned on observations (if y occurred, what did FC predict?)



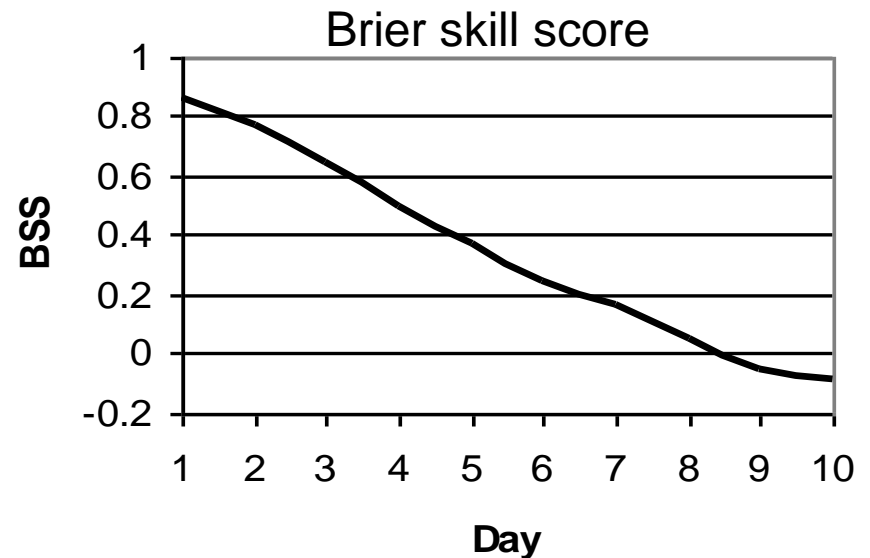
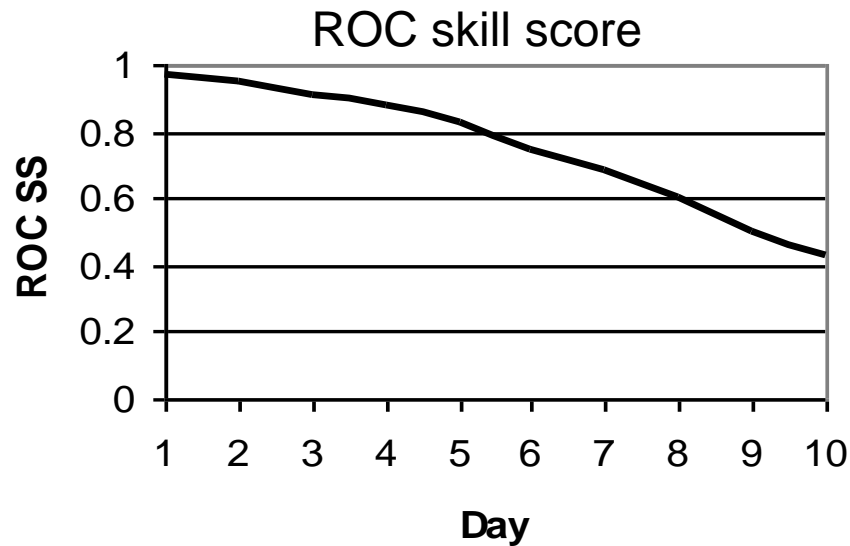
ROCSS vs. BSS

$$ROCSS = 2A - 1$$

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

- ROCSS or BSS > 0 indicate skilful forecast system

Northern Extra-Tropics 500 hPa anomalies > 2σ (spring 2002)



Richardson, 2005



Summary I

- A forecast has skill if it predicts the observed conditions well according to some objective or subjective criteria
- To evaluate a forecast system we need to look at a (large) number of forecast – observation pairs
- Different scores measure different characteristics of the forecast system: Reliability / Resolution, Brier Score (BSS), ROC,...
- Perception of usefulness of ensemble may vary with score used
- It is important to understand the behaviour of different scores and choose appropriately