Verification of Probability Forecasts

Beth Ebert
Bureau of Meteorology Research Centre (BMRC)
Melbourne, Australia

3rd International Verification Methods Workshop, 29 January – 2 February 2007

Topics

- Verification philosophy for probability forecasts
- Measuring bias
  - Reliability diagram
- Measuring total error
  - Brier score
  - Sources of error – reliability, resolution, uncertainty
- Measuring potential skill
  - Relative Operating Characteristic (ROC)

- - - - - - - - - - - - if time… - - - - - - - - - - - - - - - -

- Measuring accuracy
  - Ranked probability score
- Measuring value
  - Relative value diagram

Question:

If the forecast was for 80% chance of rain and it rained was this a good forecast?

- ☐ Yes
- ☐ No
- ☐ Don't know

Question:

If the forecast was for 10% chance of rain and it rained was this a good forecast?

Question:

Would you dare to make a prediction of 100% probability of a tornado?

- ☐ Yes
- ☐ No
- ☐ Don't know

Measuring quality of probability forecasts

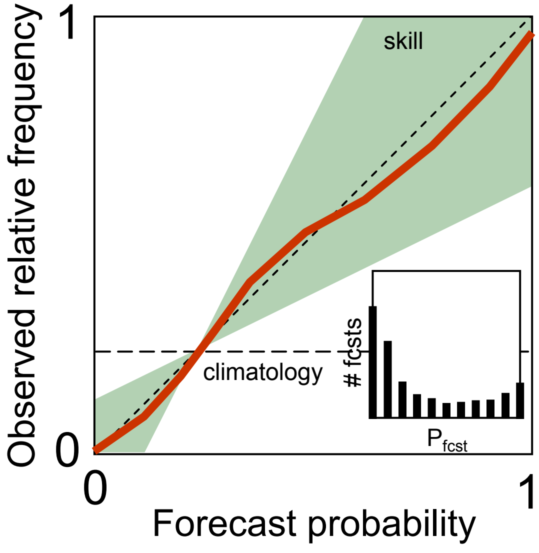
- An individual probabilistic forecast is neither completely correct or completely incorrect*

  * unless it is exactly 0% or exactly 100%

- Need to look at a large number of forecasts and observations to evaluate:
  - ***Reliability*** – can I trust the probabilities to mean what they say they mean?
  - ***Discrimination*** – how well do the forecasts distinguish between events and non-events?
  - ***Skill*** – are the forecasts better than chance or climatology?

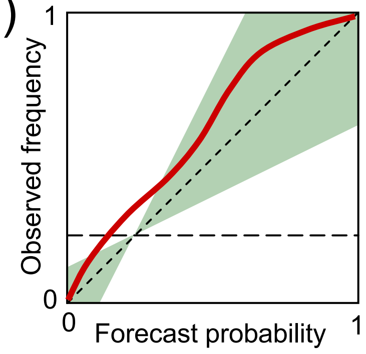
Reliability – are the forecasts unbiased?

- Measure agreement between predicted probabilities and observed frequencies
- If the forecast system is ***reliable***, then whenever the forecast probability of an event occurring is *P*, that event should occur a fraction *P* of the time.
  - For each probability category plot the frequency of observed occurrence



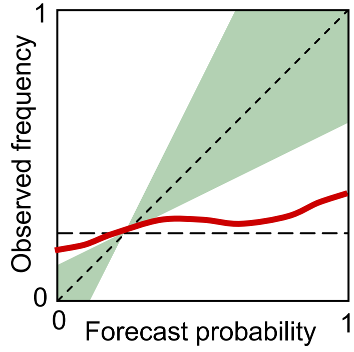
Interpretation of reliability diagrams

(a)



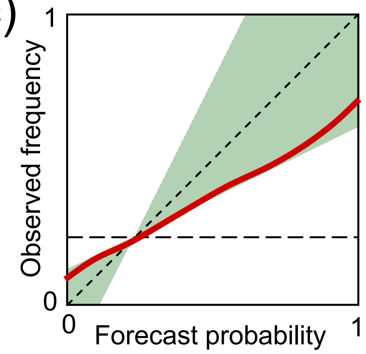
under-confident

(b)



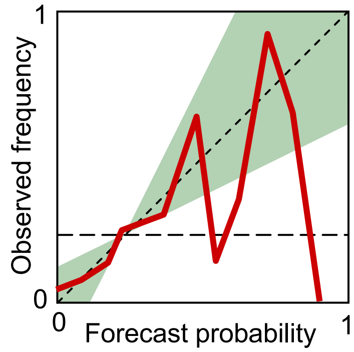
no resolution

(c)



insufficient resolution

(d)



probably under-sampled

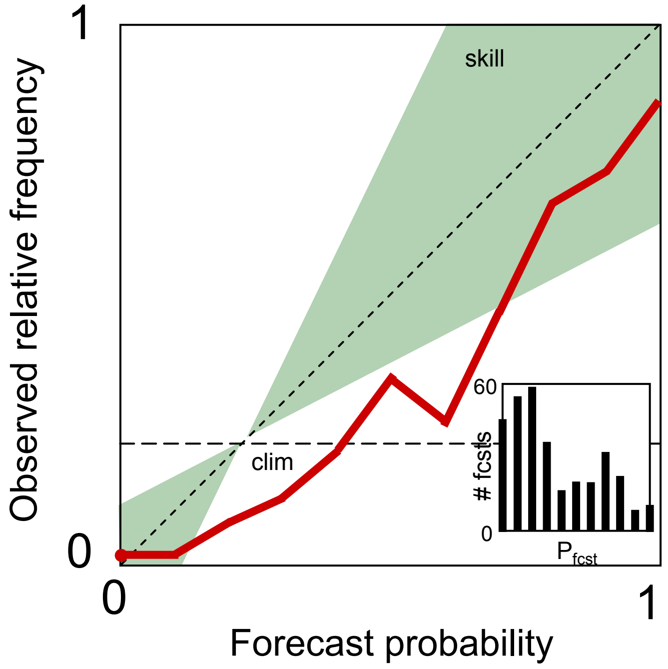
- The reliability diagram is conditioned on the forecasts (i.e., *given that X was predicted, what was the outcome?*)
- Gives information on the real meaning of the forecast.

Tampere (Finland) POP data

Date 2003	Observed rain	24h forecast POP	48h forecast POP
Jan 1	no	0.3	0.1
Jan 2	no	0.1	0.1
Jan 3	no	0.1	0.2
Jan 4	no	0.2	0.2
Jan 5	no	0.2	0.2
. . .	. . .	. . .	. . .
Dec 27	yes	0.8	0.8
Dec 28	yes	1.0	0.5
Dec 29	yes	0.9	0.9
Dec 30	no	0.1	0.3
Dec 31	no	0.1	0.1

Tampere (Finland) 24h POP summary

Forecast probability	# fcsts	# observed occurrences	Obs. relative frequency
0.0	46	1	0.02
0.1	55	1	0.02
0.2	59	5	0.08
0.3	41	5	0.12
0.4	19	4	0.21
0.5	22	8	0.36
0.6	22	6	0.27
0.7	34	16	0.47
0.8	24	16	0.67
0.9	11	8	0.73
1.0	13	11	0.85
Total	346	81	0.23 Sample climatology



Steps for making reliability diagram

1. For each forecast probability category count the number of observed occurrences
2. Compute the observed relative frequency in each category *k*

   obs. relative frequency$_k$ = obs. occurrences$_k$ / num. forecasts$_k$

3. Plot observed relative frequency vs forecast probability
4. Plot sample climatology ("no resolution" line)

   sample climatology = obs. occurrences / num. forecasts

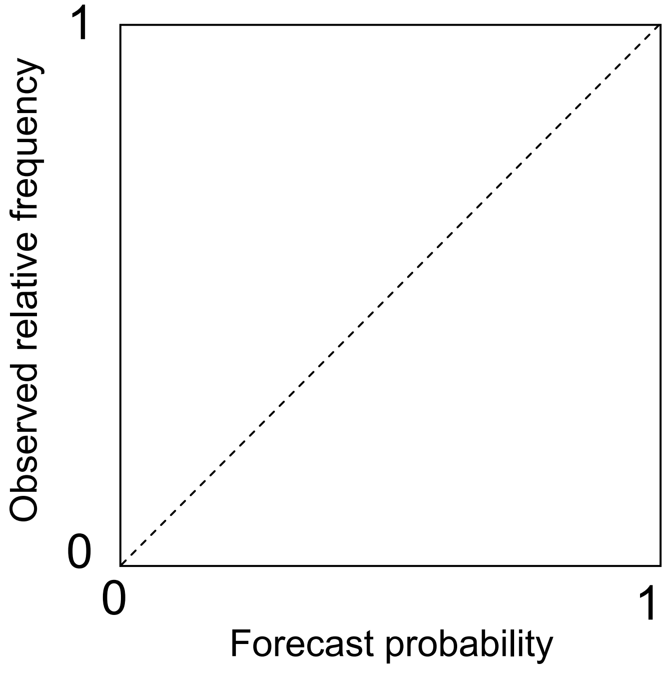
5. Plot "no-skill" line halfway between climatology and perfect reliability (diagonal) lines
6. Plot forecast frequency separately to show forecast sharpness

Tampere reliability for 48h forecasts

Forecast probability	# fcsts	# observed occurrences	Obs. relative frequency
0.0	31	1	
0.1	53	5	
0.2	67	7	
0.3	39	7	
0.4	38	12	
0.5	16	5	
0.6	26	8	
0.7	30	14	
0.8	31	15	
0.9	8	6	
1.0	7	6	

Total

Sample climatology

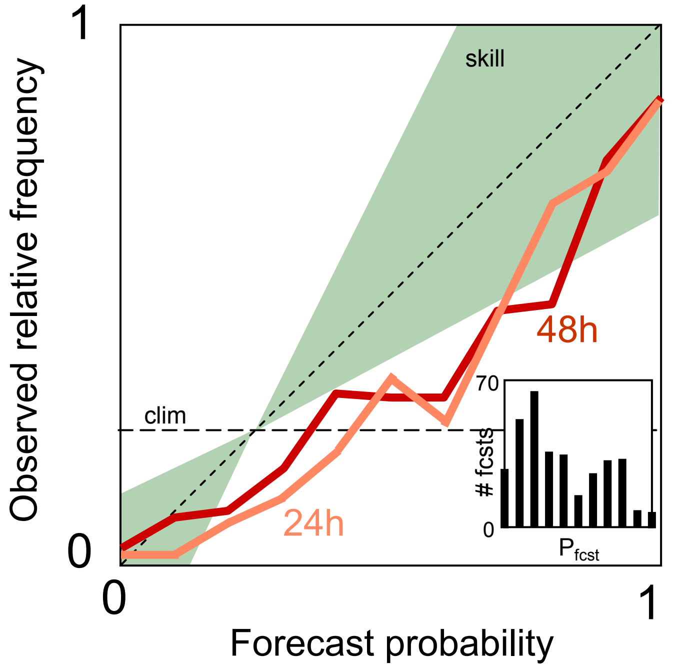


Tampere reliability for 48h forecasts

Forecast probability	# fcsts	# observed occurrences	Obs. relative frequency
0.0	31	1	0.03
0.1	53	5	0.09
0.2	67	7	0.10
0.3	39	7	0.18
0.4	38	12	0.32
0.5	16	5	0.31
0.6	26	8	0.31
0.7	30	14	0.47
0.8	31	15	0.48
0.9	8	6	0.75
1.0	7	6	0.86

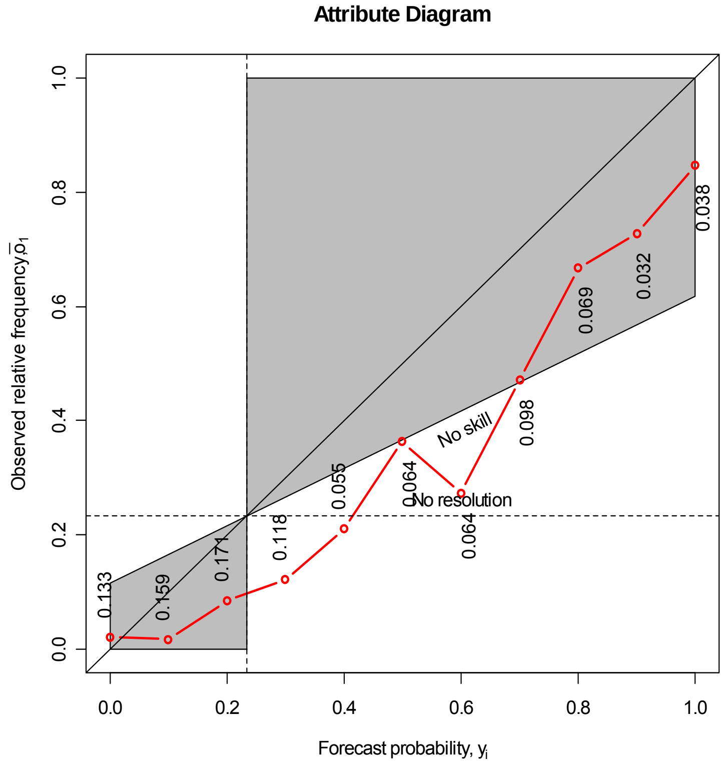
Total 346 86 0.25

Sample climatology



Reliability diagrams in R

```
library(verification)
source("read_tampere_pop.r")
A <- verify(d$obs_rain, d$p24_rain, bins=FALSE)
attribute(A)
```



Brier score – what is the probability error?

- Familiar mean square error measures accuracy of continuous variables

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - x_i)^2$$

- Brier (probability) score measures mean squared error in probability space

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

p_i = forecast probability
o_i = observed occurrence (0 or 1)

- Brier skill score measures ***skill*** relative to a reference forecast (usually climatology)

$$BSS = -\frac{BS - BS_{ref}}{BS_{ref}}$$

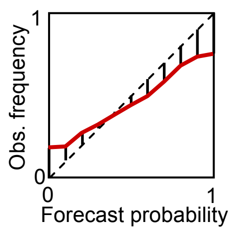
Components of probability error

The Brier score can be decomposed into 3 terms (for *K* probability classes and *N* samples):

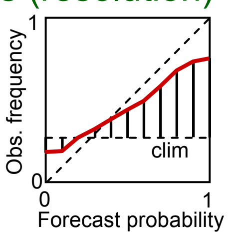
$$BS = \frac{1}{N}\sum_{k=1}^{K} n_k (p_k - \overline{o}_k)^2 - \frac{1}{N}\sum_{k=1}^{K} n_k (\overline{o}_k - \overline{o})^2 + \overline{o}(1-\overline{o})$$

reliability — resolution — uncertainty

Reliability: Measures weighted (by forecast frequency) error of reliability curve – indicates the degree to which forecast probabilities can be taken at face value (reliability)



Resolution: Measures the distance between the observed relative frequency and climatological frequency – indicates the degree to which the forecast can separate different situations (resolution)



Uncertainty: Measures the variability of the observations – indicates the degree to which situations are climatologic-ally easy or difficult to predict.

Has nothing to do with forecast quality! Use the Brier skill score to overcome this problem.

Steps for computing Brier (skill) score

1. For each forecast-observation pair compute the difference between the forecast probability p_i and observed occurrence o_i,
2. Compute the mean squared value of these differences

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

3. Compute the mean observed occurrence $\bar{o}$ (sample climatology)
4. Compute the reference Brier score using the sample climatology as the forecast (or use long-term climatology if available)

$$BS_{ref} = \frac{1}{N}\sum_{i=1}^{N}(p_i - \bar{o})^2$$

5. Compute the skill score $BSS = -\frac{BS - BS_{ref}}{BS_{ref}}$

Brier score and components in R

```
library(verification)
source("read_tampere_pop.r")
A <- verify(d$obs_rain, d$p24_rain, bins=FALSE)
summary(A)
```

```
The forecasts are probabilistic, the observations
       are binary.
Sample baseline calculated from observations.
Brier Score (BS)             =  0.1445
Brier Score - Baseline       =  0.1793
Skill Score                  =  0.1942
Reliability                  =  0.02536
Resolution                   =  0.06017
Uncertainty                  =  0.1793
```

Brier score for heavy rain vs all rain

```
H <- verify(d$obs_heavy, d$p24_heavy, bins=FALSE)
summary(H)
```

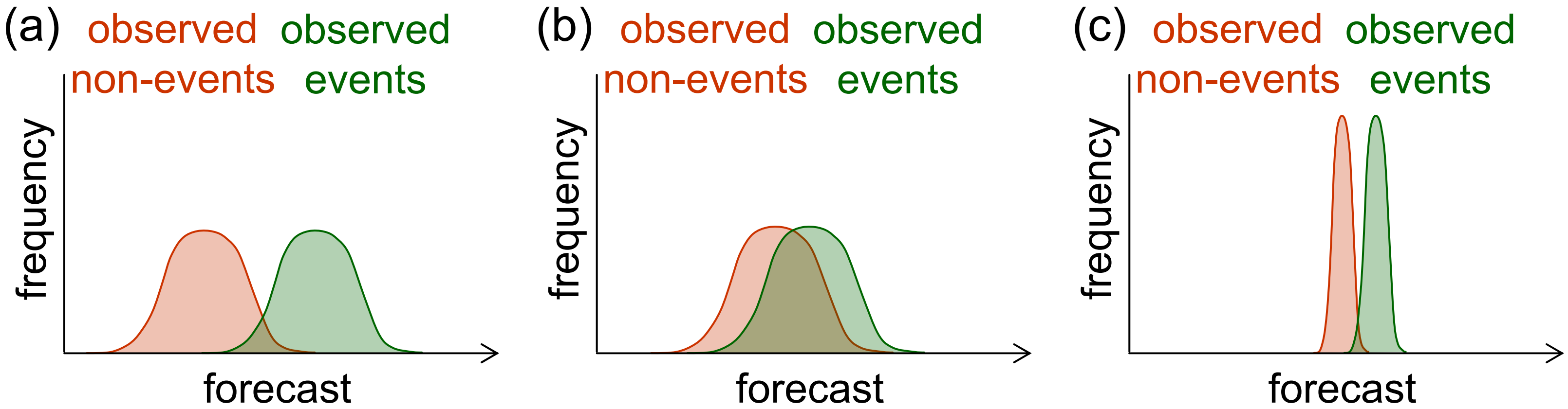
		Heavy rain	All rain
Brier Score (BS)	=	**0.03746**	**0.1445**
Brier Score - Baseline	=	**0.05446**	**0.1793**
Skill Score	=	**0.3122**	**0.1942**
Reliability	=	**0.003398**	**0.02536**
Resolution	=	**0.0204**	**0.06017**
Uncertainty	=	**0.05446**	**0.1793**

Q: What's going on?

Brier score is sensitive to the climatological frequency of an event: the more rare an event, the easier it is to get a good BS without having any real skill .

Discrimination

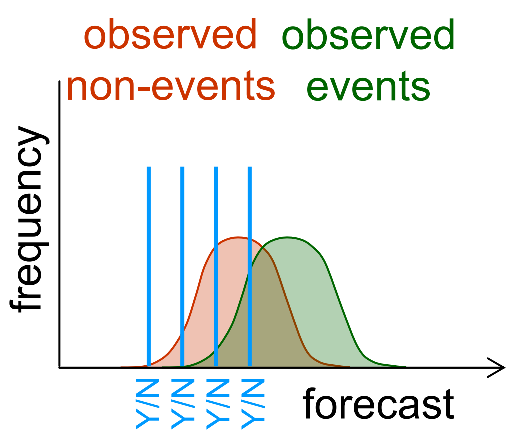
- Good forecasts should discriminate between events and non-events

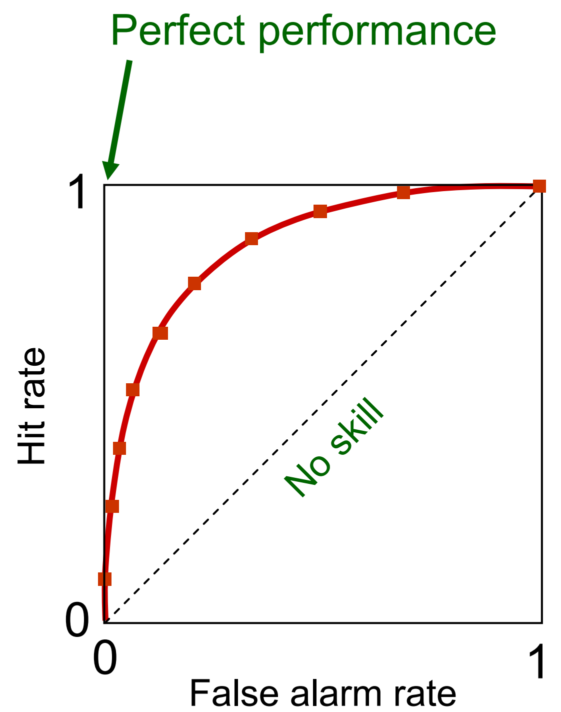


Good discrimination | Poor discrimination | Good discrimination

Measuring discrimination using ROC

- Measure success using Relative Operating Characteristic (ROC)
  - Plot the hit rate against the false alarm rate using increasing probability thresholds to make the yes/no decision



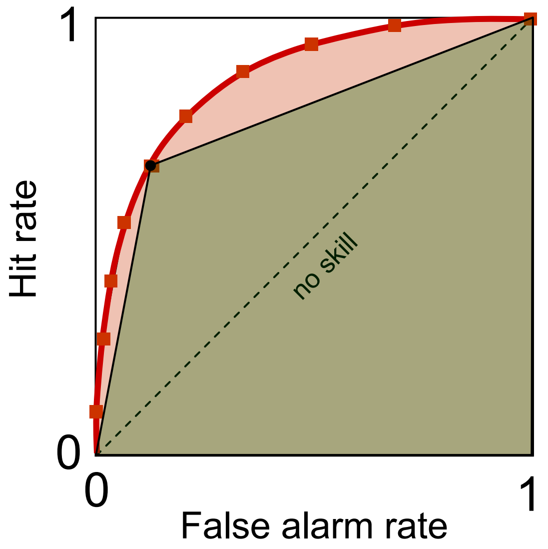


ROC area – a popular summary measure

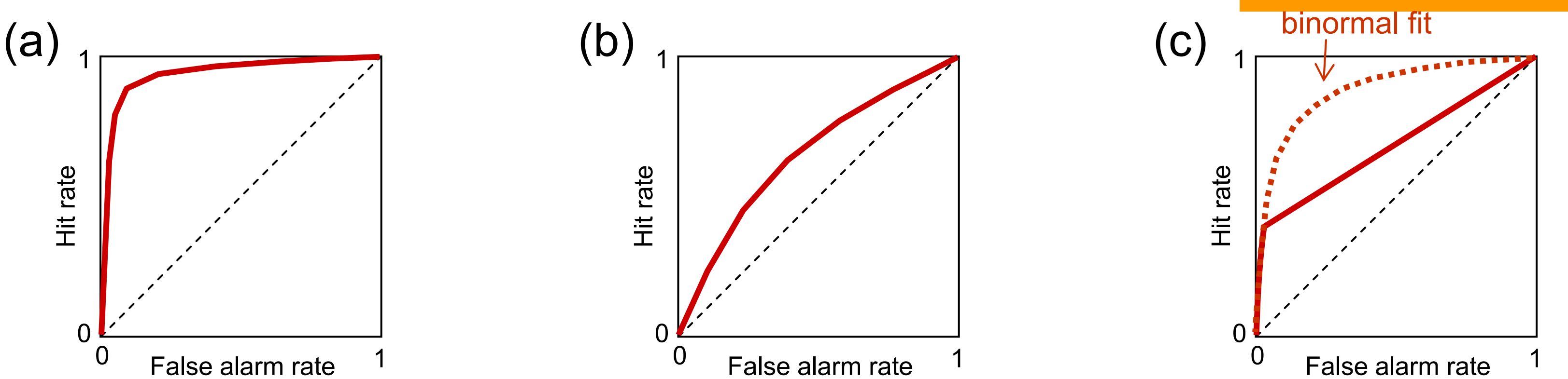
- ROC curve is independent of forecast bias – is like "potential skill"
- Area under curve ("ROC area") is a useful summary measure of forecast skill
  - Perfect: ROC area = 1
  - No skill: ROC area = 0.5
  - ROC skill score

$$ROCS = 2\,(\text{ROC area} - 0.5)$$

= KSS for deterministic forecast



Interpretation of ROC curves



high resolution (high potential skill)

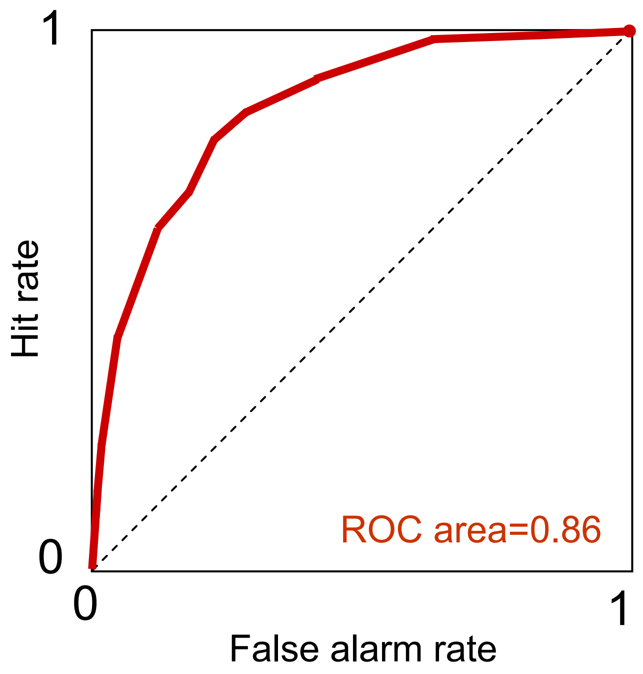
low resolution (low potential skill)

too few probability categories

- The ROC is conditioned on the observations (i.e., *given that Y occurred, how did the forecast perform?*)
- ROC is a good companion to reliability plot, which is conditioned on the forecasts (i.e., *given that X was predicted, what was the outcome?*)

Tampere ROC for 24h forecasts

Forecast probability	Hits	Misses	False alarms	Corr. non-events	Hit rate	False alarm rate
0.0	81	0	265	0	1.00	1.00
0.1	80	1	220	45	0.99	0.83
0.2	79	2	166	99	0.98	0.63
0.3	74	7	112	153	0.91	0.42
0.4	69	12	76	189	0.85	0.29
0.5	65	16	61	204	0.80	0.23
0.6	57	24	47	218	0.70	0.18
0.7	51	30	31	234	0.63	0.12
0.8	35	46	13	252	0.43	0.05
0.9	19	62	5	260	0.23	0.02
1.0	11	70	2	263	0.14	0.01



Steps for making ROC diagram

1. For each forecast probability category count the number of hits, misses, false alarms, and correct non-events
2. Compute the hit rate (probability of detection) and false alarm rate (probability of false detection) in each category *k*

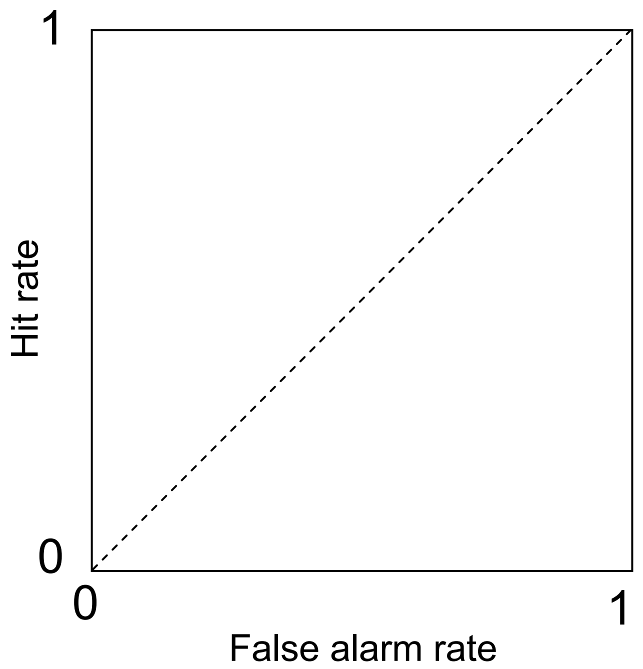
   $$\text{hit rate}_k = \text{hits}_k \;/\; (\text{hits}_k + \text{misses}_k)$$

   $$\text{false alarm rate}_k = \text{false alarms}_k \;/\; (\text{false alarms}_k + \text{correct non-events}_k)$$

3. Plot hit rate vs false alarm rate
4. ROC area is the integrated area under the ROC curve

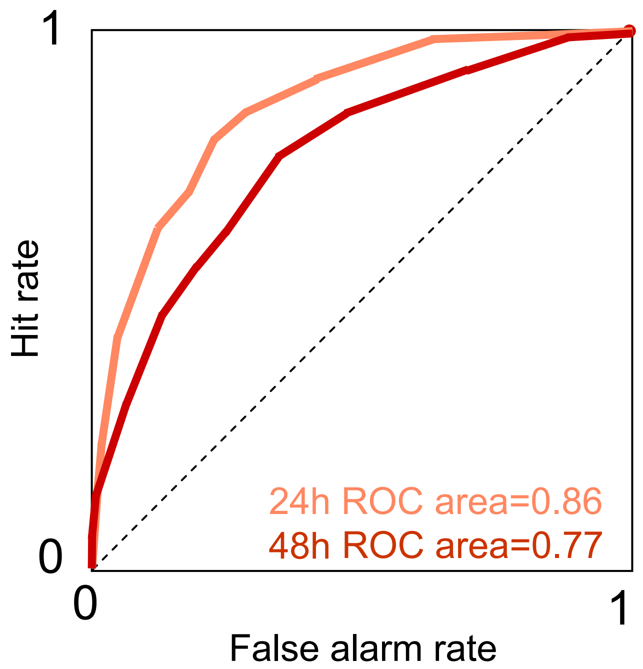
Tampere ROC for 48h forecasts

Forecast probability	Hits	Misses	False alarms	Corr. non-events	Hit rate	False alarm rate
0.0	86	0	260	0		
0.1	85	1	230	30		
0.2	80	6	182	78		
0.3	73	13	122	138		
0.4	66	20	90	170		
0.5	54	32	64	196		
0.6	49	37	53	207		
0.7	41	45	35	225		
0.8	27	59	19	241		
0.9	12	74	3	257		
1.0	6	80	1	259		



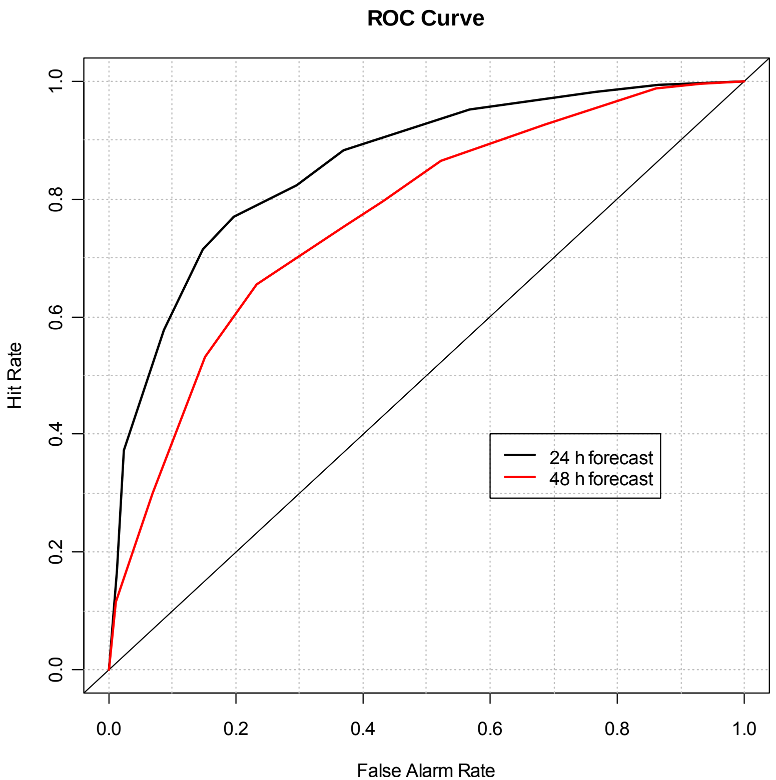
Tampere ROC for 48h forecasts

Forecast probability	Hits	Misses	False alarms	Corr. non-events	Hit rate	False alarm rate
0.0	86	0	260	0	1.00	1.00
0.1	85	1	230	30	0.99	0.89
0.2	80	6	182	78	0.93	0.70
0.3	73	13	122	138	0.85	0.47
0.4	66	20	90	170	0.77	0.35
0.5	54	32	64	196	0.63	0.25
0.6	49	37	53	207	0.57	0.20
0.7	41	45	35	225	0.48	0.13
0.8	27	59	19	241	0.31	0.07
0.9	12	74	3	257	0.14	0.01
1.0	6	80	1	259	0.07	0.00



ROC diagrams in R

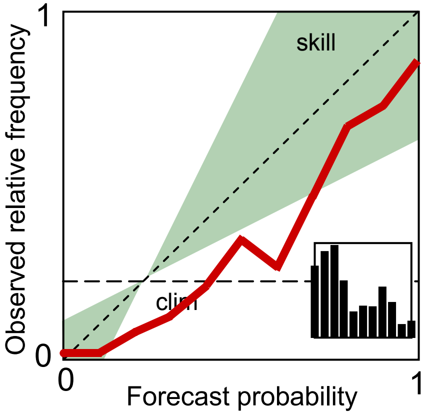
```
library(verification)
source("read_tampere_pop.r")
A <- verify(d$obs_rain, d$p24_rain, bins=FALSE)
roc.plot(A, legend=TRUE)

roc.plot(A, CI=TRUE)

roc.plot(A, binormal=TRUE,
    plot="both", legend=TRUE,
    show.thres=FALSE)


B <- verify(d$obs_rain,
    d$p48_rain, bins=FALSE)
roc.plot(A, plot.thres=NULL)
lines.roc(B, col=2, lwd=2)
leg.txt <- c("24 h forecast",
    "48 h forecast")
legend(0.6, 0.4, leg.txt,
    col=c(1,2), lwd=2)
```



Putting it all together…

Tampere POP forecasts

- Reliability diagram
  - measures bias



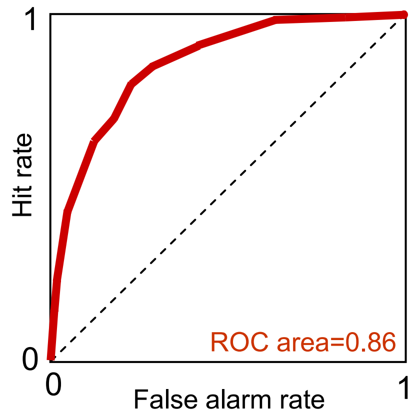
high bias (over-confident), better than climatology only for *P* near 0 or 1

- Brier score
  - measures probability error

Brier Score (BS) = 0.1445
Brier Skill Score = 0.1942

skilled compared to climatology

- ROC
  - measures discrimination (potential skill)

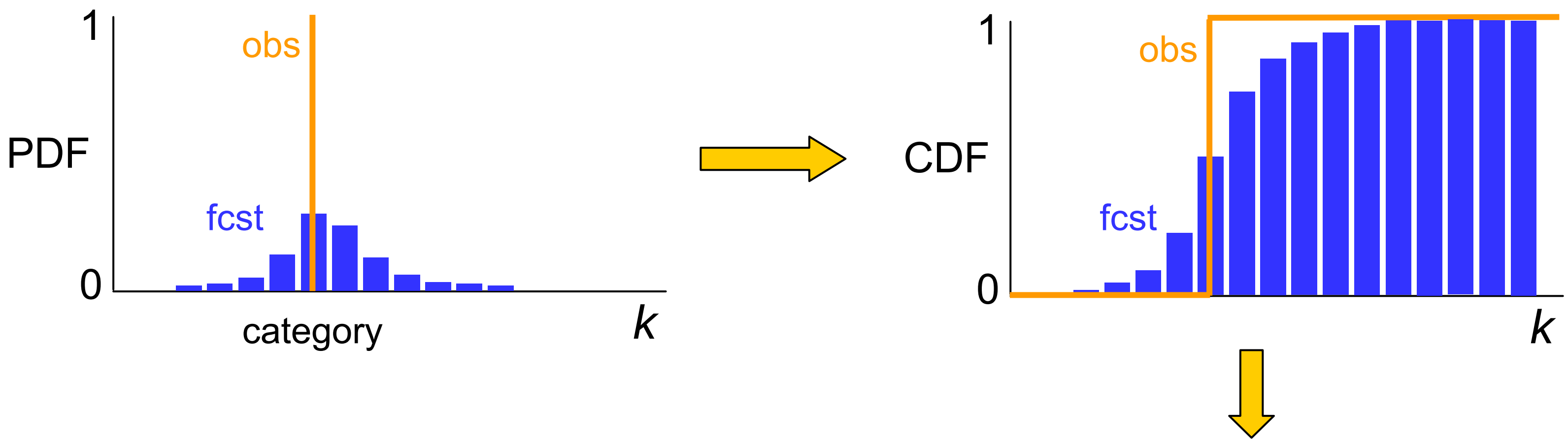


good discrimination→ good potential skill

… more probability verification …

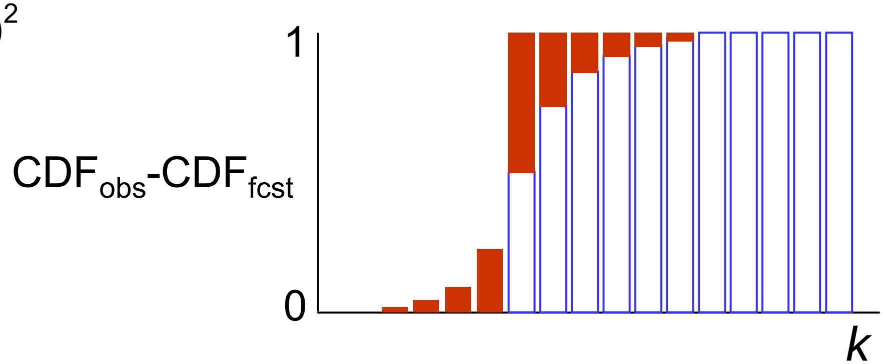
Ranked probability score – how accurate are the probability forecasts?

Measures the squared difference in probability space when there are multiple probability categories



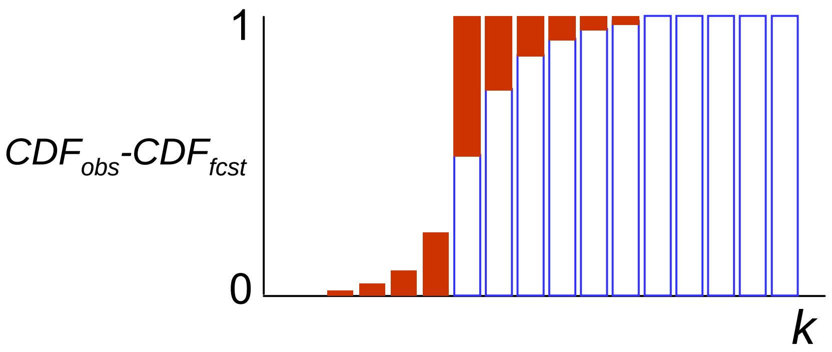
$$RPS = \frac{1}{K-1} \sum_{k=1}^{K} (CDF_{fcst,k} - CDF_{obs,k})^2$$

for *K* probability classes



Characteristics of RPS

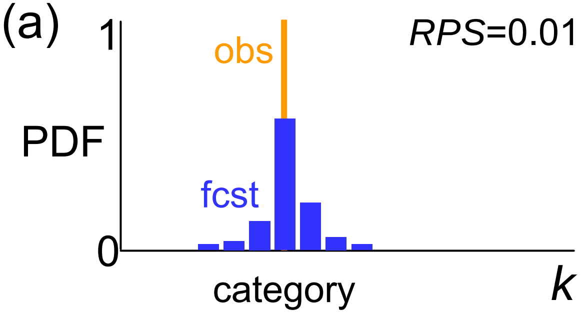
$$RPS = \frac{1}{K-1}\sum_{k=1}^{K}(CDF_{fcst,k} - CDF_{obs,k})^2$$



- Takes into account the ordered nature of the predicted variable (for example, temperature going from low to high values)
- Emphasizes accuracy by penalizing "near misses" less than larger errors
- Rewards sharp forecast if it is accurate
- Perfect score: 0
- RPS skill score w.r.t. climatology: $RPSS = 1 - \frac{RPS}{RPS_{c\,lim}}$

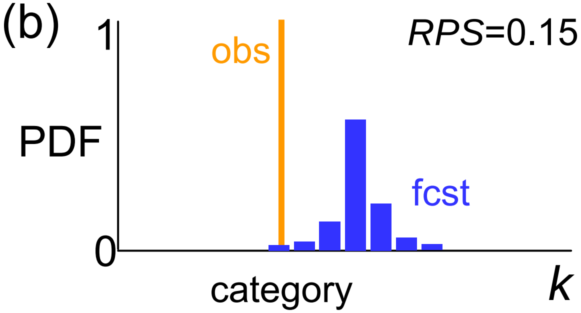
Interpretation of RPS

(a)



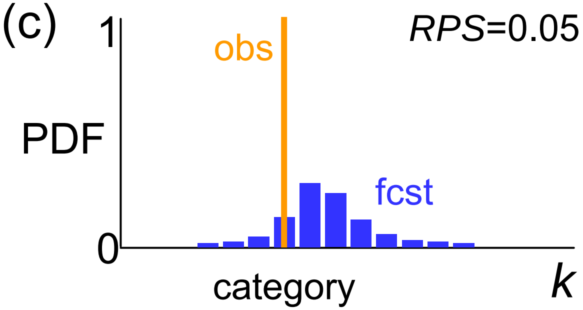
forecast is sharp and accurate
→ small RPS (good)

(b)



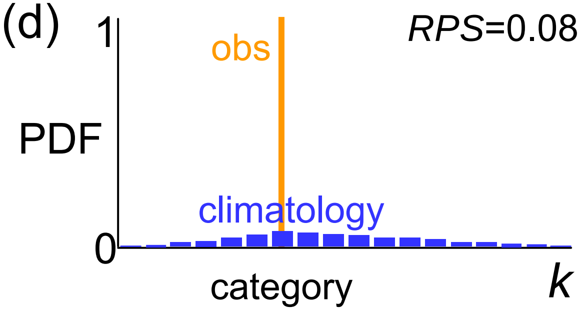
forecast is sharp but biased
→ large RPS (bad)

(c)



forecast not very sharp and slightly biased
→ moderate RPS

(d)



forecast is accurate but not sharp
→ moderate RPS

Q: Which forecasts are skilled with respect to climatology?

Tampere 24h POP data

Date 2003	Observed rain (mm)	p_1 = POP 0-0.2 mm (category 1)	p_2 = POP 0.3-4.4 mm (category 2)	p_3 = POP 4.5+ mm (category 3)
Jan 1	0.0	0.7	0.3	0.0
Jan 2	0.0	0.9	0.1	0.0
Jan 3	0.0	0.9	0.1	0.0
Jan 4	0.0	0.8	0.2	0.0
Jan 5	0.0	0.8	0.2	0.0
Jan 6	0.0	0.9	0.1	0.0
Jan 7	1.1	0.6	0.4	0.0
Jan 8	0.9	0.3	0.4	0.3
Jan9	0.0	0.3	0.4	0.3
Jan 10	0.0	NA	NA	NA
Jan 11	2.2	NA	NA	NA
Jan 12	0.0	0.8	0.2	0.0
Jan 13	1.2	0.8	0.2	0.0
Jan 14	6.0	0.0	0.4	0.6
Jan 15	2.3	0.3	0.7	0.0
. . .	. . .	. . .	. . .	. . .

Steps for computing RPS

1. For each forecast-observation pair:

   a. Assign the observation to its appropriate category $k_{obs.}$ The cumulative density function CDF_{obs} is either 0 or 1:

   $$CDF_{obs,k} = \begin{cases} 0 & k < k_{obs} \\ 1 & k \geq k_{obs} \end{cases}$$

   b. From the categorical probability forecast $P = [p_1, p_2, \ldots, p_K]$ compute the cumulative density function for every category k as

   $$CDF_{fcst,k} = \sum_{j=1}^{k} p_j$$

   c. Compute the RPS as $RPS = \frac{1}{K-1}\sum_{k=1}^{K}(CDF_{fcst,k} - CDF_{obs,k})^2$

2. Average the RPS over all forecast-observation pairs

Tampere 24h POP data

Categories
1 ≤ 0.2 mm
2 0.3 - 4.4 mm
3 ≥ 4.5 mm

Date 2003	Observed rain (mm)	p_1	p_2	p_3	Observed category k_{obs}	$CDF_{obs,k}$ k=1,2,3	$CDF_{fcst,k}$ k=1,2,3	*RPS*
Jan 1	0.0	0.7	0.3	0.0	1	1, 1, 1	0.7, 1, 1	0.045
Jan 2	0.0	0.9	0.1	0.0	1	1, 1, 1	0.9, 1, 1	0.005
Jan 3	0.0	0.9	0.1	0.0	1	1, 1, 1	0.9, 1, 1	0.005
Jan 4	0.0	0.8	0.2	0.0	1	1, 1, 1	0.8, 1, 1	0.020
Jan 5	0.0	0.8	0.2	0.0	1	1, 1, 1	0.8, 1, 1	0.020
Jan 6	0.0	0.9	0.1	0.0	1	1, 1, 1	0.9, 1, 1	0.005
Jan 7	1.1	0.6	0.4	0.0	2	0, 1, 1	0.6, 1, 1	0.180
Jan 8	0.9	0.3	0.4	0.3	2	0, 1, 1	0.3, 0.7, 1	0.090
Jan9	0.0	0.3	0.4	0.3	1	1, 1, 1	0.3, 0.7, 1	0.290
Jan 10	0.0	NA	NA	NA	1	1, 1, 1	NA	NA
Jan 11	2.2	NA	NA	NA	2	0, 1, 1	NA	NA
Jan 12	0.0	0.8	0.2	0.0	1	1, 1, 1	0.8, 1, 1	0.020
Jan 13	1.2	0.8	0.2	0.0	2	0, 1, 1	0.8, 1, 1	0.320
Jan 14	6.0	0.0	0.4	0.6	3	0, 0, 1	0, 0.4, 1	0.080
Jan 15	2.3	0.3	0.7	0.0	2	0, 1, 1	0.3, 1, 1	0.045
. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .

15-day RPS = 0.087

Tampere 48h POP data

Categories

1 ≤ 0.2 mm
2 0.3 - 4.4 mm
3 ≥ 4.5 mm

Date 2003	Observed rain (mm)	p_1	p_2	p_3	Observed category k_{obs}	$CDF_{obs,k}$ k=1,2,3	$CDF_{fcst,k}$ k=1,2,3	*RPS*
Jan 1	0.0	0.9	0.1	0.0				
Jan 2	0.0	0.9	0.1	0.0				
Jan 3	0.0	0.8	0.1	0.1				
Jan 4	0.0	0.8	0.1	0.1				
Jan 5	0.0	0.8	0.2	0.0				
Jan 6	0.0	0.8	0.2	0.0				
Jan 7	1.1	0.8	0.2	0.0				
Jan 8	0.9	0.7	0.3	0.0				
Jan9	0.0	0.4	0.4	0.2				
Jan 10	0.0	0.8	0.1	0.1				
Jan 11	2.2	NA	NA	NA				
Jan 12	0.0	NA	NA	NA				
Jan 13	1.2	0.6	0.4	0.0				
Jan 14	6.0	0.1	0.5	0.4				
Jan 15	2.3	0.2	0.6	0.2				
. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .

15-day RPS =

Tampere 48h RPS

<u>Categories</u>

1 ≤ 0.2 mm
2 0.3 - 4.4 mm
3 ≥ 4.5 mm

Date 2003	Observed rain (mm)	p_1	p_2	p_3	Observed category k_{obs}	$CDF_{obs,k}$ k=1,2,3	$CDF_{fcst,k}$ k=1,2,3	*RPS*
Jan 1	0.0	0.9	0.1	0.0	1	1,1,1	0.9, 1, 1	0.005
Jan 2	0.0	0.9	0.1	0.0	1	1,1,1	0.9, 1, 1	0.005
Jan 3	0.0	0.8	0.1	0.1	1	1,1,1	0.8, 0.9, 1	0.025
Jan 4	0.0	0.8	0.1	0.1	1	1,1,1	0.8, 0.9, 1	0.025
Jan 5	0.0	0.8	0.2	0.0	1	1,1,1	0.8, 1, 1	0.020
Jan 6	0.0	0.8	0.2	0.0	1	1,1,1	0.8, 1, 1	0.020
Jan 7	1.1	0.8	0.2	0.0	2	0,1,1	0.8, 1, 1	0.320
Jan 8	0.9	0.7	0.3	0.0	2	0,1,1	0.7, 1, 1	0.245
Jan9	0.0	0.4	0.4	0.2	1	1,1,1	0.4, 0.8, 1	0.200
Jan 10	0.0	0.8	0.1	0.1	1	1,1,1	0.8, 0.9, 1	0.025
Jan 11	2.2	NA	NA	NA	2	0,1,1	NA	NA
Jan 12	0.0	NA	NA	NA	1	1,1,1	NA	NA
Jan 13	1.2	0.6	0.4	0.0	2	0,1,1	0.6, 1, 1	0.180
Jan 14	6.0	0.1	0.5	0.4	3	0,0,1	0.1, 0.6, 1	0.185
Jan 15	2.3	0.2	0.6	0.2	2	0,1,1	0.2, 0.8, 1	0.040
. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .

15-day RPS
=0.100

Ranked probability (skill) score in R

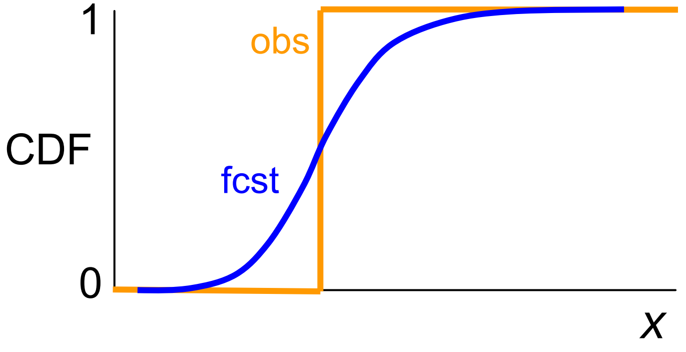
```
library(verification)
source("read_tampere_pop.r")
# Make vector of observed categories
obscat <- d$obs_norain + d$obs_light*2 + d$obs_heavy*3
# Make Nx3 array of category probabilities
pvec   <- cbind(d$p24_norain, d$p24_light, d$p24_heavy)
rps(obscat, pvec)
```

```
$rps
[1] 0.0909682

$rpss
[1] 0.2217009

$rps.clim
[1] 0.1168808
```

Continuous ranked probability score

Continuous ranked probability score (CRPS) measures the difference between the forecast and observed CDFs

$$CRPS = \int_{-\infty}^{\infty} \left(P_{fcst}(x) - P_{obs}(x) \right)^2 dx$$



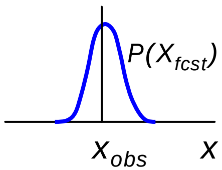
- Same as Brier score integrated over all possible threshold values
- Same as Mean Absolute Error for deterministic forecasts
- Advantages:
  - sensitive to whole range of values of the parameter of interest
  - does not depend on predefined classes
  - easy to interpret
  - has dimensions of the observed variable
- Rewards small spread (sharpness) if the forecast is accurate
- Perfect score: 0

Verifying individual events

Debate as to whether or not this is a good idea…

- Forecasters and other users often want to know the quality of a forecast for a particular event
- We cannot meaningfully verify a single probability forecast
  - If it rains when the PoP was 30% was that a good forecast?

- ... but we can compare a probability distribution to a single observation
  - Want the forecast to be accurate (close to the observed), and sharp (not too much spread)
  - **This approach implicitly assumes that the weather is *predictable* and the uncertainty comes from the *forecast system***
  - Best used at short time ranges and/or large spatial scales

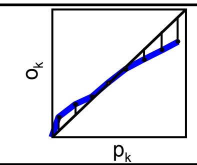
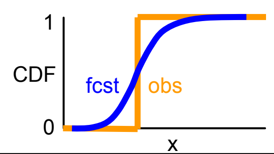
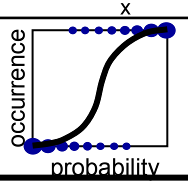


- Methods for individual or collections of forecasts
  - (Continuous) Ranked Probability Score
  - Wilson (MWR, 1999) score
  - Ignorance

Conveying forecast quality to users

Forecasters and other users are ~comfortable with standard verification measures for deterministic forecasts

Are there similar easy-to-understand measures for probabilistic forecasts?

Deterministic	**Probabilistic** (suggestions)	**Visual aid**
Mean bias	Reliability term of BS $\frac{1}{N}\sum_{k=1}^{K} n_k (p_k - \bar{o}_k)^2$	
RMS error	Brier score (square root) $\sqrt{BS} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(p_i - o_i)^2}$	
Mean absolute error	CRPS $\int (P_{fcst}(x) - P_{obs}(x))^2 dx$	
Correlation	R^2 for logistic regression	

Relative value score

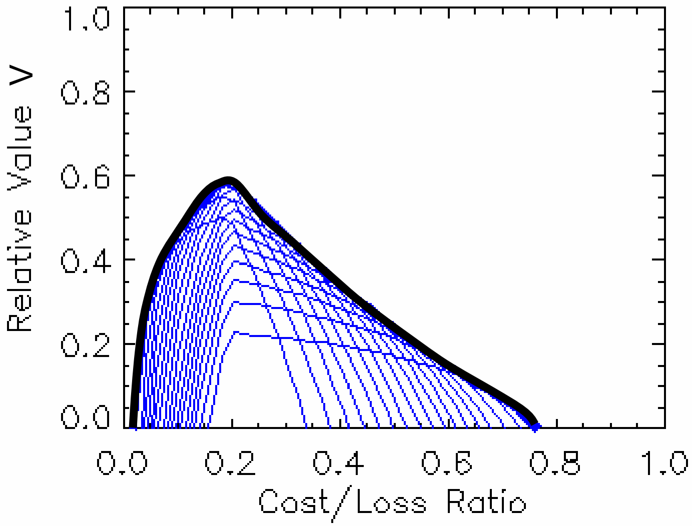
Measures the relative improvement in economic value as a function of the cost/loss ratio C/L for taking action based on a forecast as opposed to climatology

$$V = (1-F) - \left(\frac{1-C/L}{C/L}\right)\left(\frac{\bar{o}}{1-\bar{o}}\right)(1-H) \qquad \text{if } C/L < \bar{o}$$

$$V = H - \left(\frac{C/L}{1-C/L}\right)\left(\frac{1-\bar{o}}{\bar{o}}\right)F \qquad \text{if } C/L > \bar{o}$$

where *H* is the hit rate and *F* is the false alarm rate

- The relative value is a skill score of expected expense, with climatology as the reference forecast.
- Range: -∞ to 1. Perfect score: 1
- Plot V vs C/L for various probability thresholds. The envelope describes the potential value for the probabilistic forecasts.

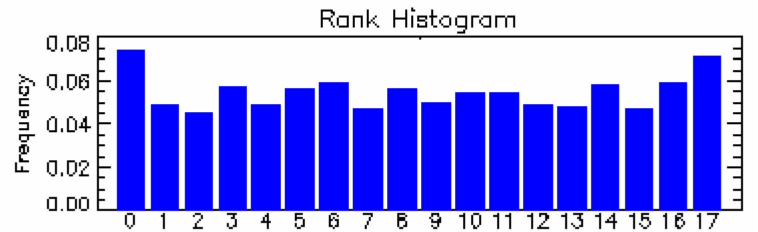


Rank histogram (Talagrand diagram)

Measures how well the ensemble spread of the forecast represents the true variability (uncertainty) of the observations

→ Count where the verifying observation falls with respect to the ensemble forecast data, which is arranged in increasing order at each grid point.

- In an ensemble with perfect spread, each member represents an equally likely scenario, so the observation is equally likely to fall between any two members.
  - Flat - ensemble spread correctly represents forecast uncertainty
  - U-shaped - ensemble spread too small, many observations falling outside the extremes of the ensemble
  - Dome-shaped - ensemble spread too large, too many observations falling near the center of the ensemble
  - Asymmetric - ensemble contains bias
- A flat rank histogram does not necessarily indicate a skilled forecast, it only measures whether the observed probability distribution is well represented by the ensemble.



Who's using what for ensemble verification?

- WMO (ensemble NWP, site maintained by JMA)
  - Brier skill score, reliability diagram, economic value, ensemble mean & spread
- Some operational centers (ensemble NWP) – web survey in 2005

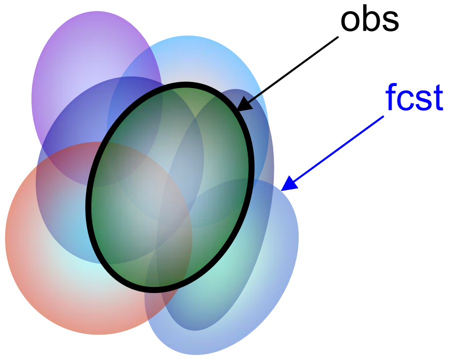
ECMWF	BSS, reliability diagram, ROC, ROC area, econ. value, spread/skill diagram
NCEP	RMSE and AC of ensemble mean, BSS, ROC area, rank histogram, RPSS, econ. value
Met Office	BSS, reliability diagram, ROC, rank histogram
BMRC	RMSE ensemble mean, BSS, reliability diagram, ROC, rank histogram, RPSS, econ. value

- DEMETER (multiple coupled-model seasonal ensemble) – see http://www.ecmwf.int/research/demeter/d/charts/verification/
  - Deterministic: anomaly correlation, mean square skill score, SD ratio
  - Probabilistic: reliability diagram, ROCS, RPSS
  - Economic value

Verifying "objects"

Significant weather events can often be viewed as 2D objects

- tropical cyclones, heavy rain events, deep low pressure centres
- objects are defined by an intensity threshold



What might the ensemble forecast look like?

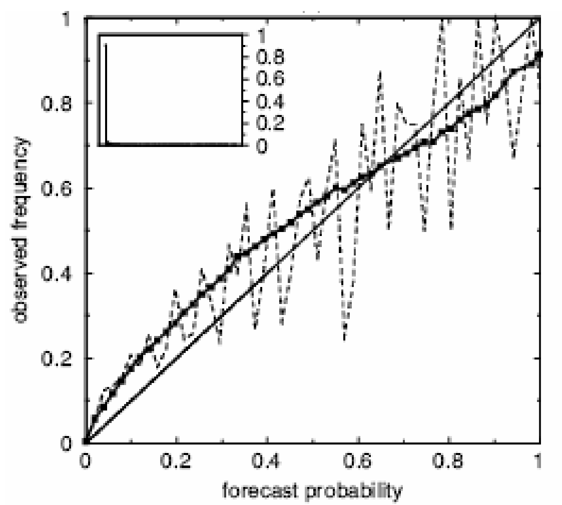
- spatial probability contour maps
- distributions of object properties
  - location, size, intensity, etc.

Strategies for verifying ensemble predictions of objects

- Verify spatial probability maps
- Verify distributions of object properties
  - many samples – use probabilistic measures
  - individual cases – CRPS, WS, IGN
- Verify ensemble mean
  - spatially averaged forecast objects
  - generated from average object properties

Sampling issues – rare events

- Rare events are often the most interesting ones!
- Coarse model resolution may not capture intensity of experienced weather
- Difficult to verify probabilities on the "tail" of the PDF
  - Too few samples to get robust statistics, especially for reliability
  - Finite number of ensemble members may not resolve tail of forecast PDF
- Forecast calibration approaches
- Atger (*QJRMS*, 2004) approach for improving robustness of verification:
  - Fit ROC for all events (incl. rare) using bi-normal model, then relate back to reliability to get *estimated* forecast quality for under-sampled categories
  - Fitted reliability also be used instead of "raw" frequencies to calibrate ensemble



Effects of observation errors

Observation errors add uncertainty to the verification results

- True forecast skill is unknown
  - $\rightarrow$ An imperfect model / ensemble may score better!
- Extra dispersion of observation PDF

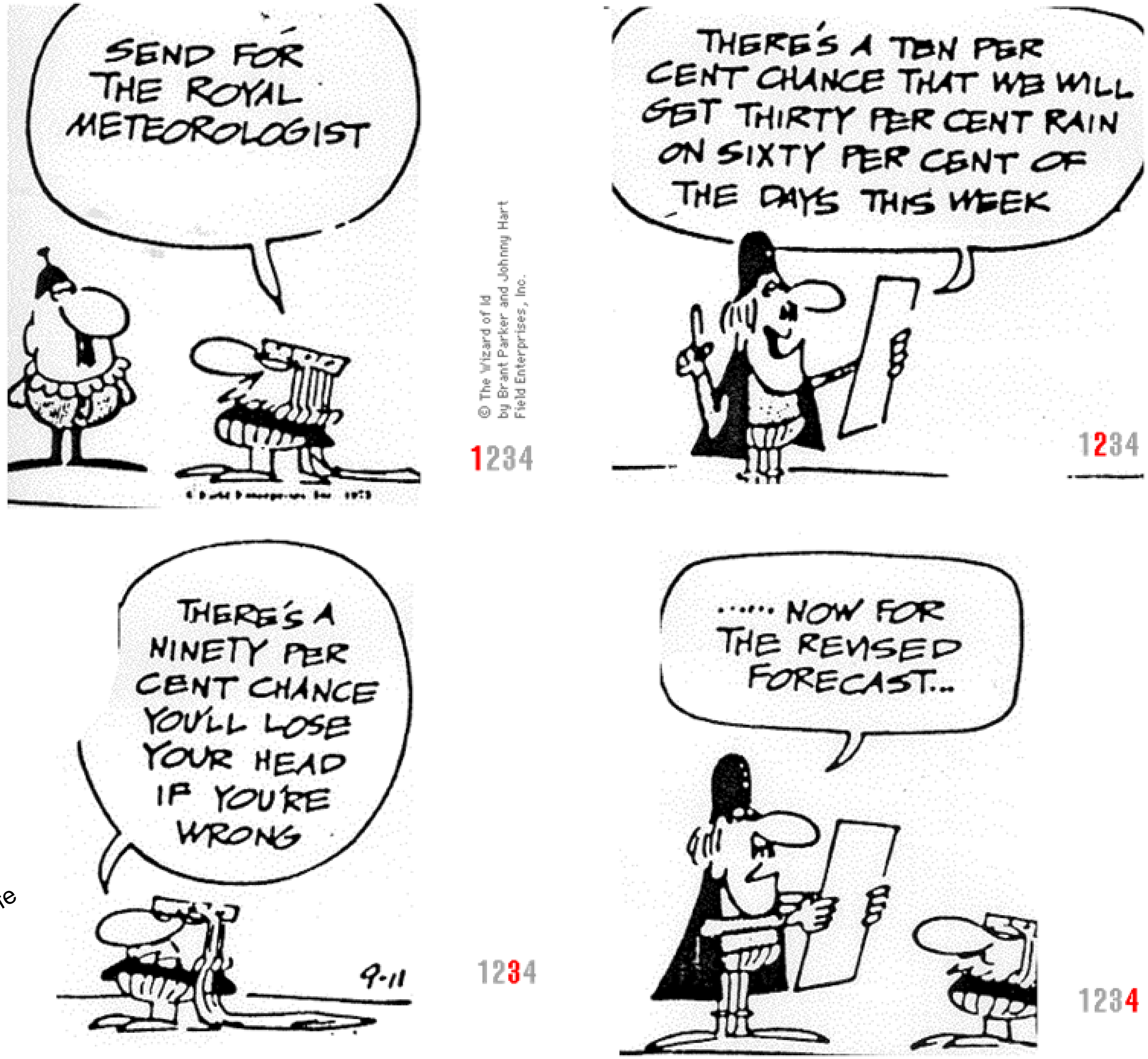
Effects on verification results

- RMSE – overestimated
- Spread – more obs outliers make ensemble look under-dispersed
  - Saetra et al (2004) compensate by adding obs error to ensemble
- Reliability – poorer
- Resolution – greater in BS decomposition, but ROC area poorer
- CRPS, WS, IGN – poorer mean values

Can we remove the effects of observation error?

- More samples helps with reliability estimates
- Error modeling – study effects of applied observation errors
- Need "gold standard" to measure actual observation errors

Not easy!



Thanks Ian Jolliffe

Thank you!