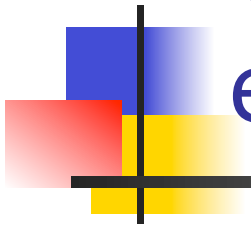
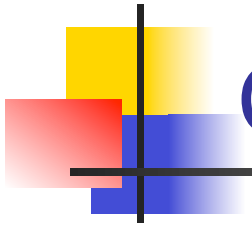


Verification of probability and ensemble forecasts

Laurence J. Wilson

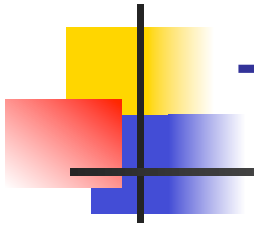
Atmospheric Science and Technology Branch

Environment Canada



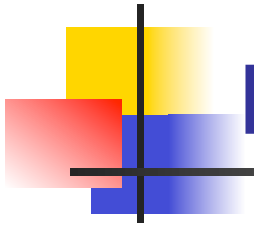
Goals of this session

- Increase understanding of scores used for probability forecast verification
  - Characteristics, strengths and weaknesses
- Know which scores to choose for different verification questions.
- Not so specifically on R – projects.



Topics

- Introduction: review of essentials of probability forecasts for verification
- Brier score: *Accuracy*
- Brier skill score: *Skill*
- Reliability Diagrams: *Reliability, resolution* and *sharpness*
  - Exercise
- *Discrimination*
  - Exercise
- Relative operating characteristic
  - Exercise
- Ensembles: The CRPS and Rank Histogram



Probability forecast

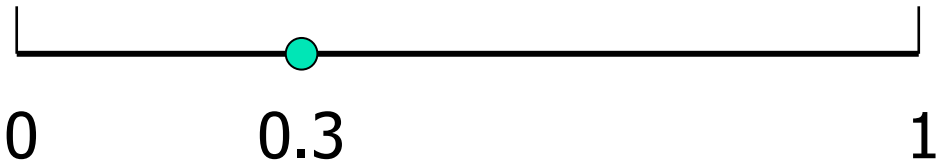
- Applies to a specific, completely defined event
  - Examples: Probability of precipitation over 6h
  - ….
- Question: What does a probability forecast "POP for Helsinki for today (6am to 6pm) is 0.95" mean?

The Brier Score

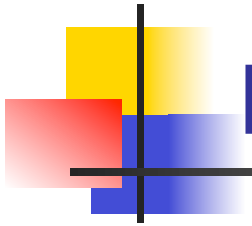
- Mean square error of a probability forecast

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

- Weights larger errors more than smaller ones



- *Sharpness*: The tendency of probability forecasts towards categorical forecasts, measured by the variance of the forecasts
  - A measure of a forecasting strategy; does not depend on obs



Brier Score

- Gives result on a single forecast, but cannot get a perfect score unless forecast categorically.
- Strictly proper
- A “summary” score – measures accuracy, summarized into one value over a dataset.
- Brier Score decomposition – components of the error

Components of probability error

The Brier score can be decomposed into 3 terms (for *K* probability classes and a sample of size *N*):

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

reliability

If for all occasions when forecast probability p_k is predicted, the observed frequency of the event is $\bar{o}_k = p_k$ then the forecast is said to be reliable. Similar to bias for a continuous variable

resolution

The ability of the forecast to distinguish situations with distinctly different frequencies of occurrence.

uncertainty

The variability of the observations. Maximized when the climatological frequency (*base rate*) =0.5

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

The presence of the uncertainty term means that Brier Scores should not be compared on different samples.

Brier Skill Score

- In the usual skill score format: proportion of improvement of accuracy over the accuracy of a standard forecast, climatology or persistence.

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

- IF the sample climatology is used, can be expressed as:

$$BSS = -\frac{\mathrm{Res} - \mathrm{Rel}}{\mathrm{Unc}}$$

Brier score and components in R

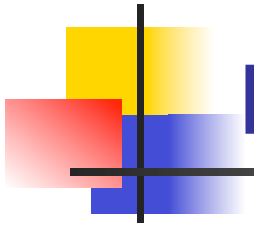
```
library(verification)

mod1 <- verify(obs = DAT$obs, pred = DAT$msc)

summary(mod1)
```

The forecasts are probabilistic, the observations are binary.
Sample baseline calculated from observations.

		1 Stn	20 Stns
Brier Score (BS)	=	0.08479	0.06956
Brier Score - Baseline	=	0.09379	0.08575
Skill Score	=	0.09597	0.1888
Reliability	=	0.01962	0.007761
Resolution	=	0.02862	0.02395
Uncertainty	=	0.09379	0.08575

Brier Score and Skill Score - Summary

- Measures accuracy and skill respectively
- "Summary" scores
- Cautions:
  - Cannot compare BS on different samples
  - BSS – take care about underlying climatology
  - BSS – Take care about small samples



Reliability Diagrams 1

- A graphical method for assessing reliability, resolution, and sharpness of a probability forecast
- Requires a fairly large dataset, because of the need to partition (bin) the sample into subsamples conditional on forecast probability
- Sometimes called “attributes” diagram.

Reliability diagram 2: How to do it

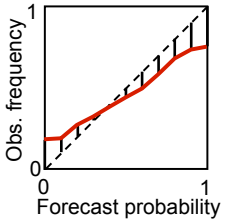
1. Decide number of categories (bins) and their distribution:
   - Depends on sample size, discreteness of forecast probabilities
   - Should be an integer fraction of ensemble size for e.g.
   - Don't all have to be the same width – within bin sample should be large enough to get a stable estimate of the observed frequency.
2. Bin the data
3. Compute the observed conditional frequency in each category (bin) *k*
   - $\textit{obs. relative frequency}_k = \textit{obs. occurrences}_k \;/\; \textit{num. forecasts}_k$
4. Plot observed frequency vs forecast probability
5. Plot sample climatology ("no resolution" line) (The sample base rate)
   - $\textit{sample climatology} = \textit{obs. occurrences} \;/\; \textit{num. forecasts}$
6. Plot "no-skill" line halfway between climatology and perfect reliability (diagonal) lines
7. Plot forecast frequency histogram to show sharpness (or plot number of events next to each point on reliability graph)

Reliability Diagram 3

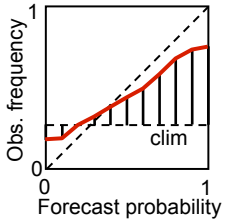
Reliability: Proximity to diagonal

Resolution: Variation about horizontal (climatology) line

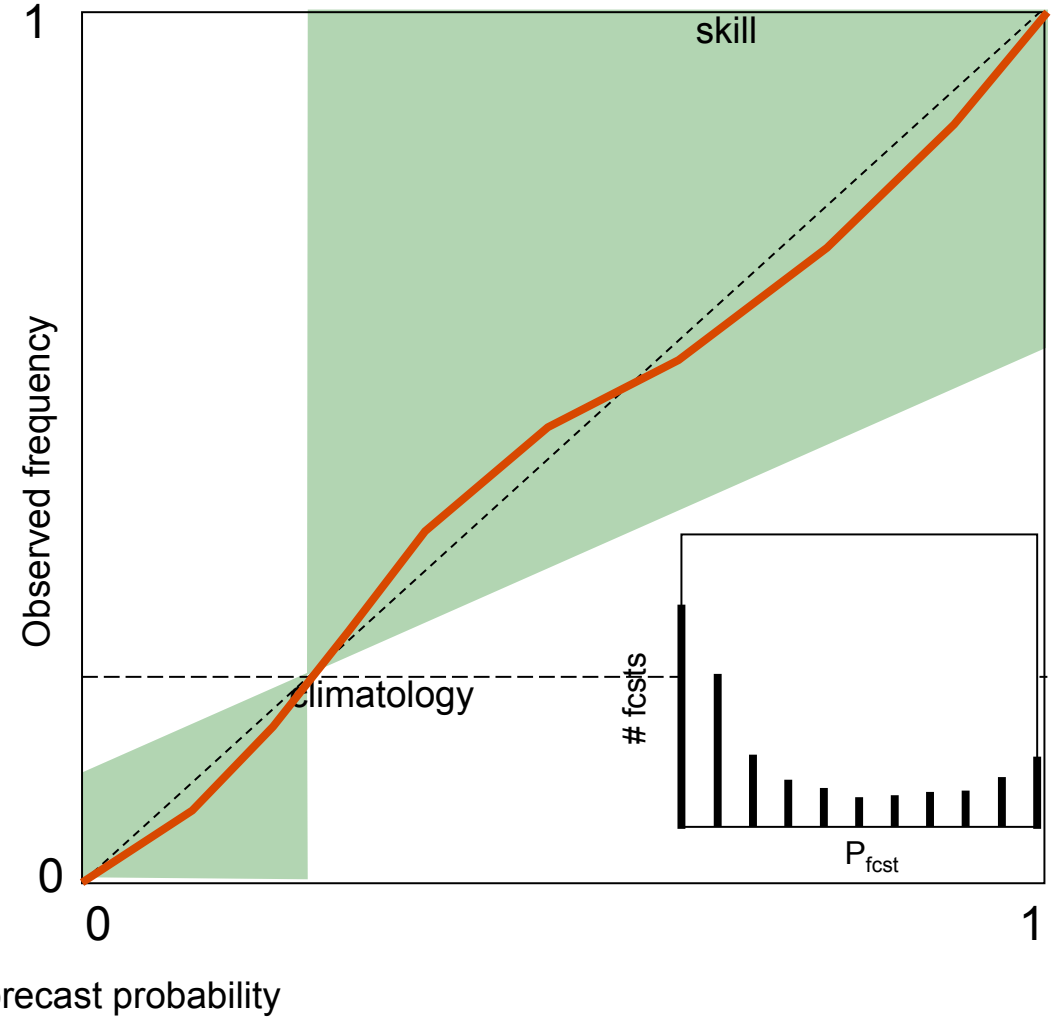
No skill line: Where reliability and resolution are equal – Brier skill score goes to 0



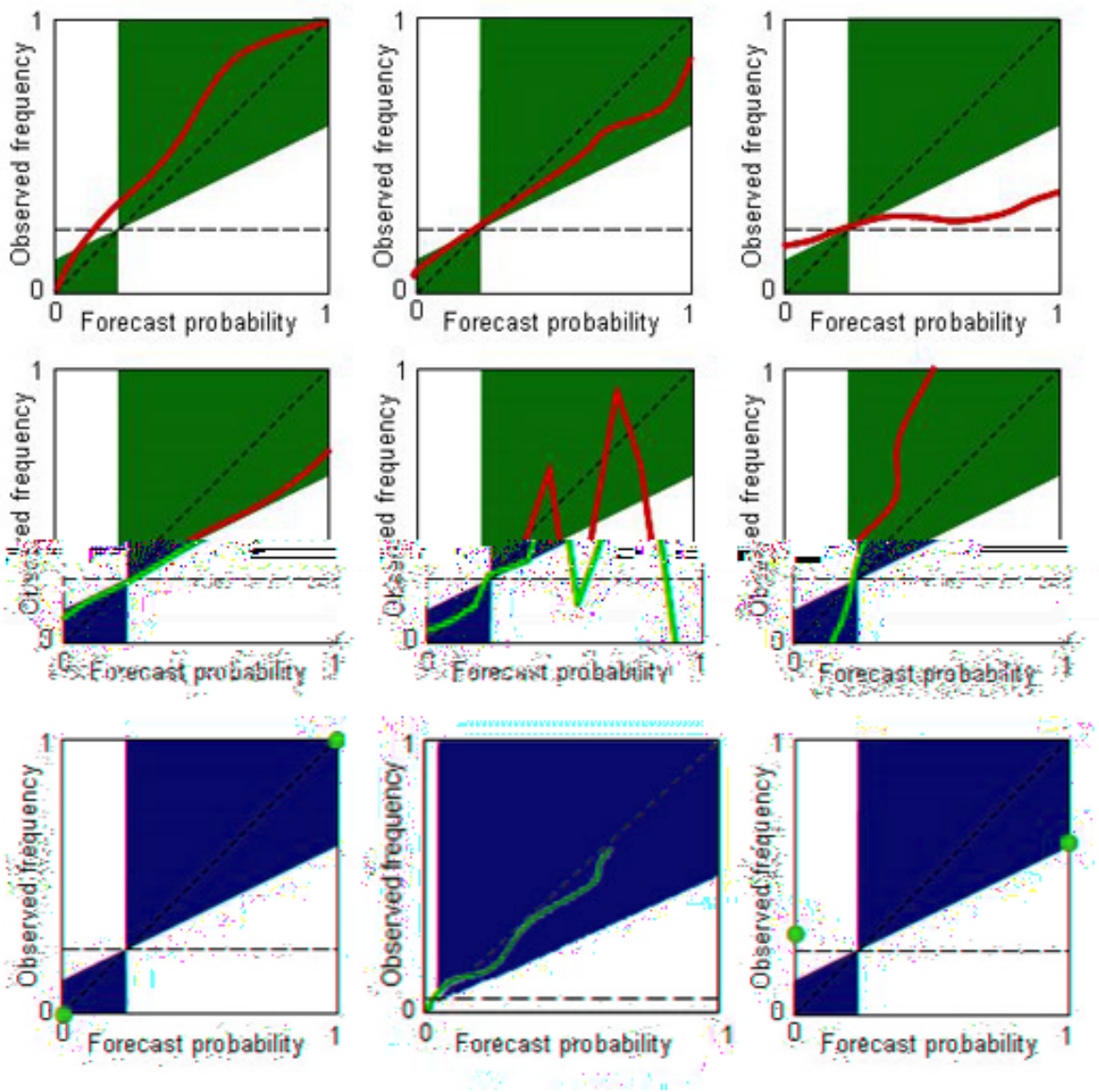
Reliability



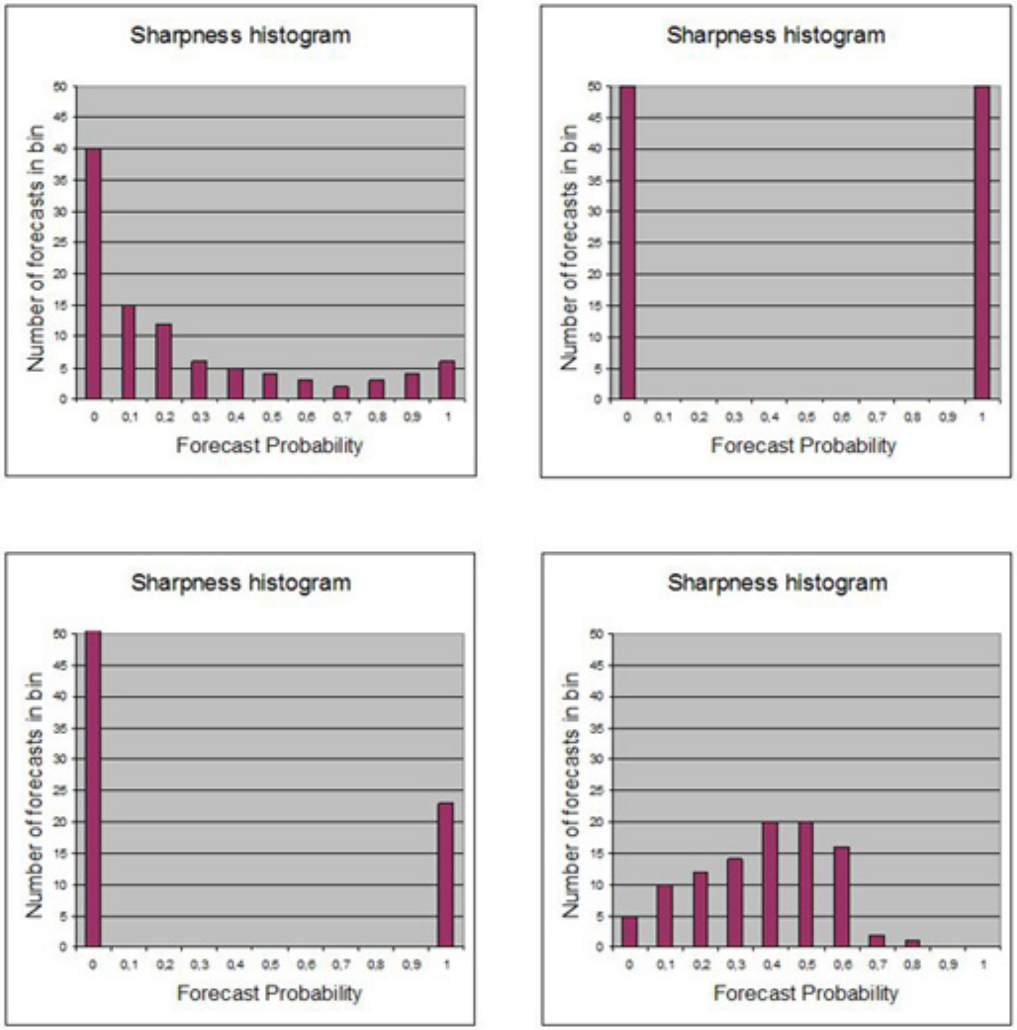
Resolution



Reliability Diagram Exercise

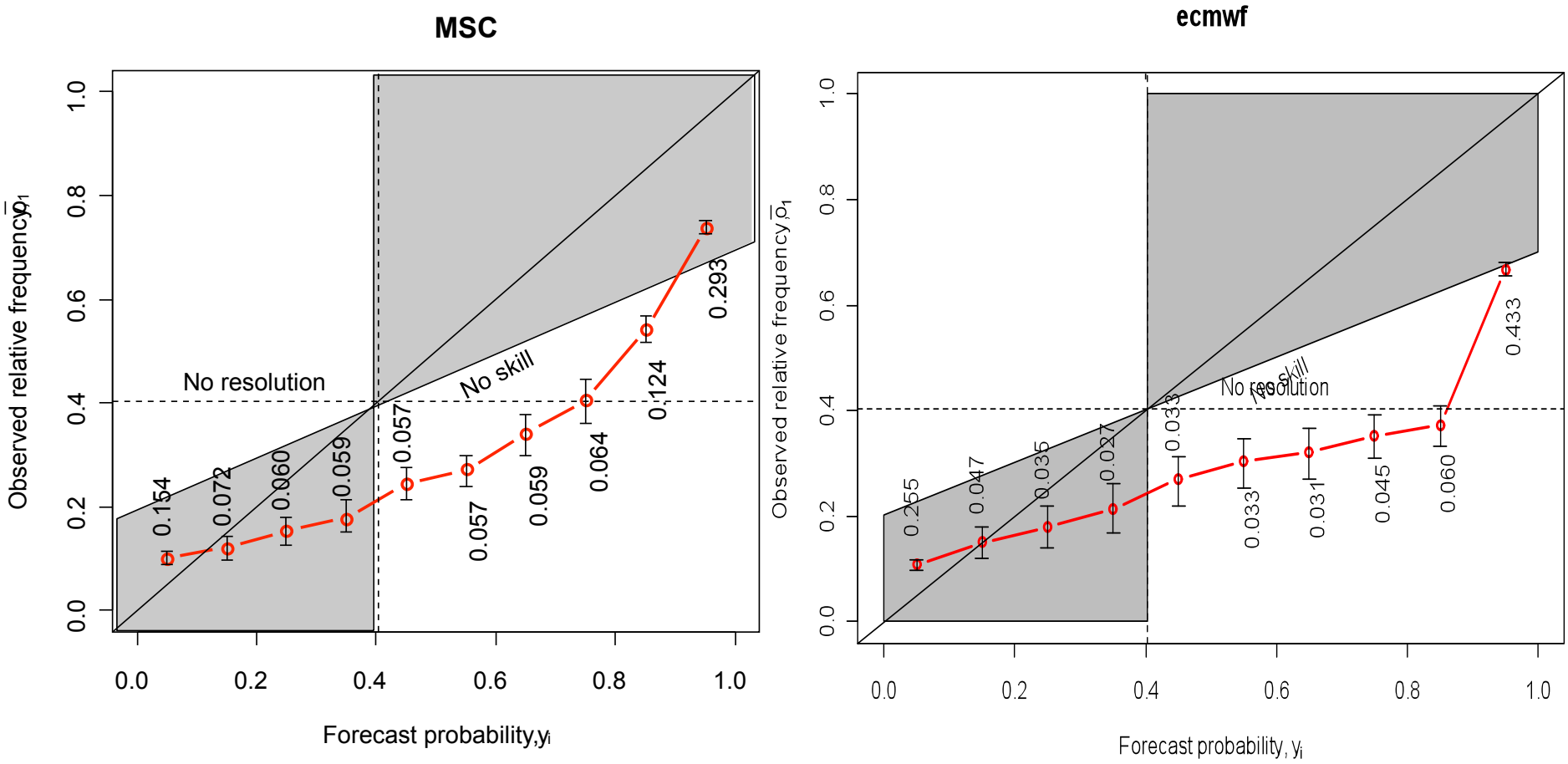


Sharpness Histogram Exercise



Reliability Diagram in R

plot(mod1, main = names(DAT)[3], CI = TRUE)



Brier score and components in R

```
library(verification)

for(i in 1:4){
        mod1 <- verify(obs = DAT$obs, pred = DAT[,1+i])
        summary(mod1)
        }
```

The forecasts are probabilistic, the observations are binary.
Sample baseline calculated from observations.

		MSC	ECMWF
Brier Score (BS)	=	0.2241	0.2442
Brier Score - Baseline	=	0.2406	0.2406
Skill Score	=	0.06858	-0.01494
Reliability	=	0.04787	0.06325
Resolution	=	0.06437	0.05965
Uncertainty	=	0.2406	0.2406

Reliability Diagram Exercise

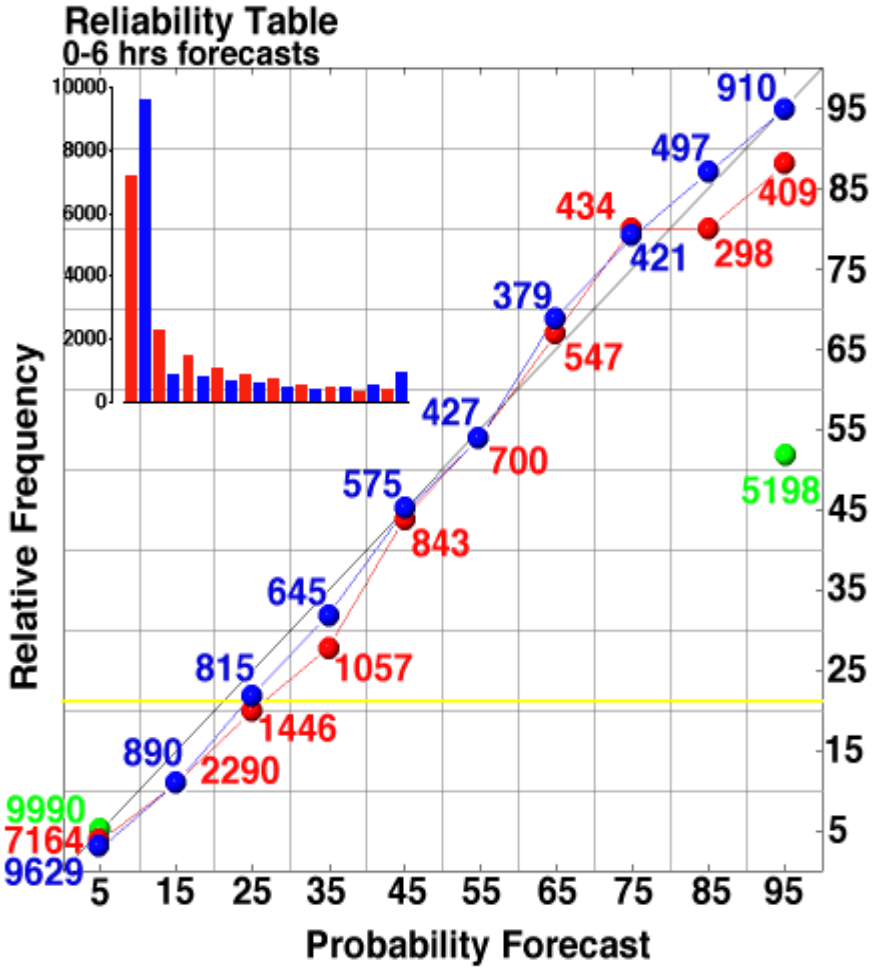
Reliability Table Exercise

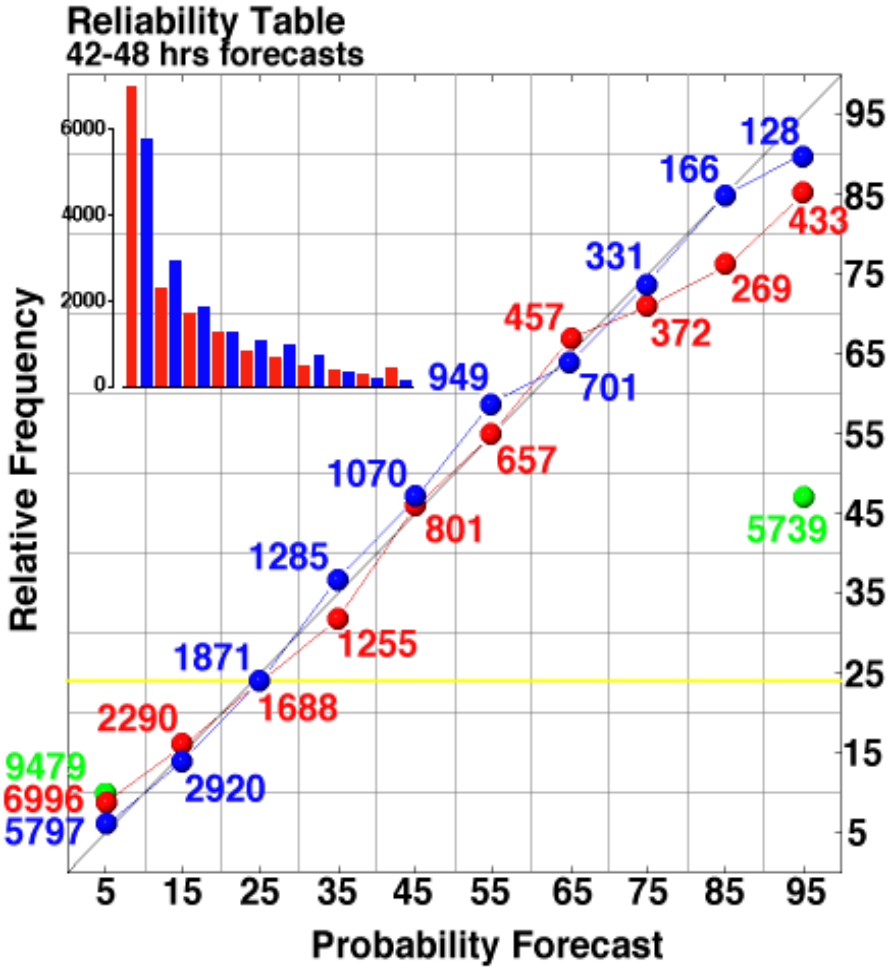
You have two reliability tables, one for 0 to 6 h POP forecasts and the other for 42 to 48 h POP forecasts (6 h periods). Forecasts are for 220 Canadian stations over a three month period, January 1, to March 31, 1999. On each graph there are two lines, representing the forecasts from two different techniques. Technique A is represented in blue and Technique B in red. In the upper left corner, the histograms indicate the number of times each of the 10 probability categories was predicted. Technique A is shown on the histograms by the blue bars and Technique B by the red bars. The frequencies of prediction of each probability category are also indicated by the numbers beside the points on the graphs. The horizontal line is the sample climatological frequency of occurrence of precipitation.

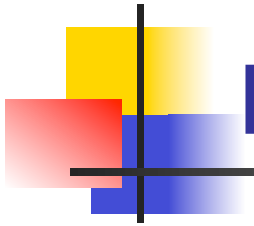
Questions:

1. Comment on the reliability of the two techniques as indicated by both tables. What does a forecast of 85% actually mean at 0 to 6 h and 42 to 48 h?
2. Which technique is sharper at 0 to 6 h? at 42 to 48 h? How do you know?
3. The two extra plotted points (in green) represent categorical forecasts of precipitation from a third technique. Comment on the reliability of this method for both forecast periods.
4. Which of the two probability forecast techniques produces the better forecasts in your opinion. Why?

Reliability Diagram Exercise

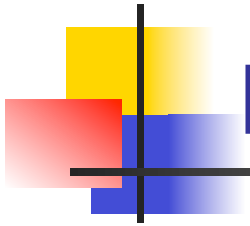






Reliability Diagrams - Summary

- Diagnostic tool
- Measures “reliability”, “resolution” and “sharpness”
- Requires “reasonably” large dataset to get useful results
- Try to ensure enough cases in each bin
- Graphical representation of Brier score components

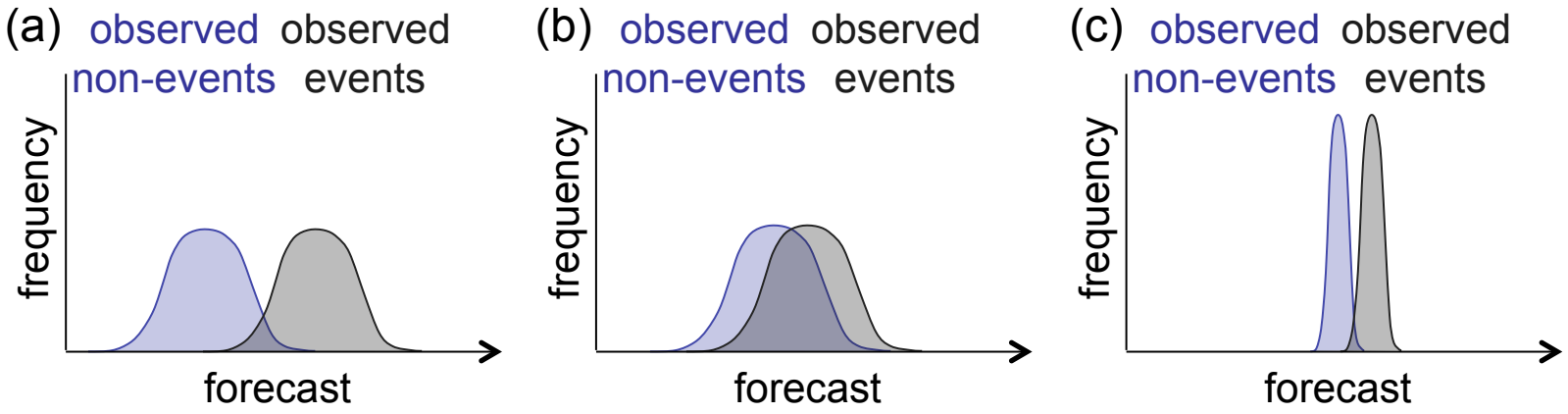


Discrimination and the ROC

- Reliability diagram – partitioning the data according to the forecast probability
- Suppose we partition according to observation – 2 categories, yes or no
- Look at distribution of forecasts separately for these two categories

Discrimination

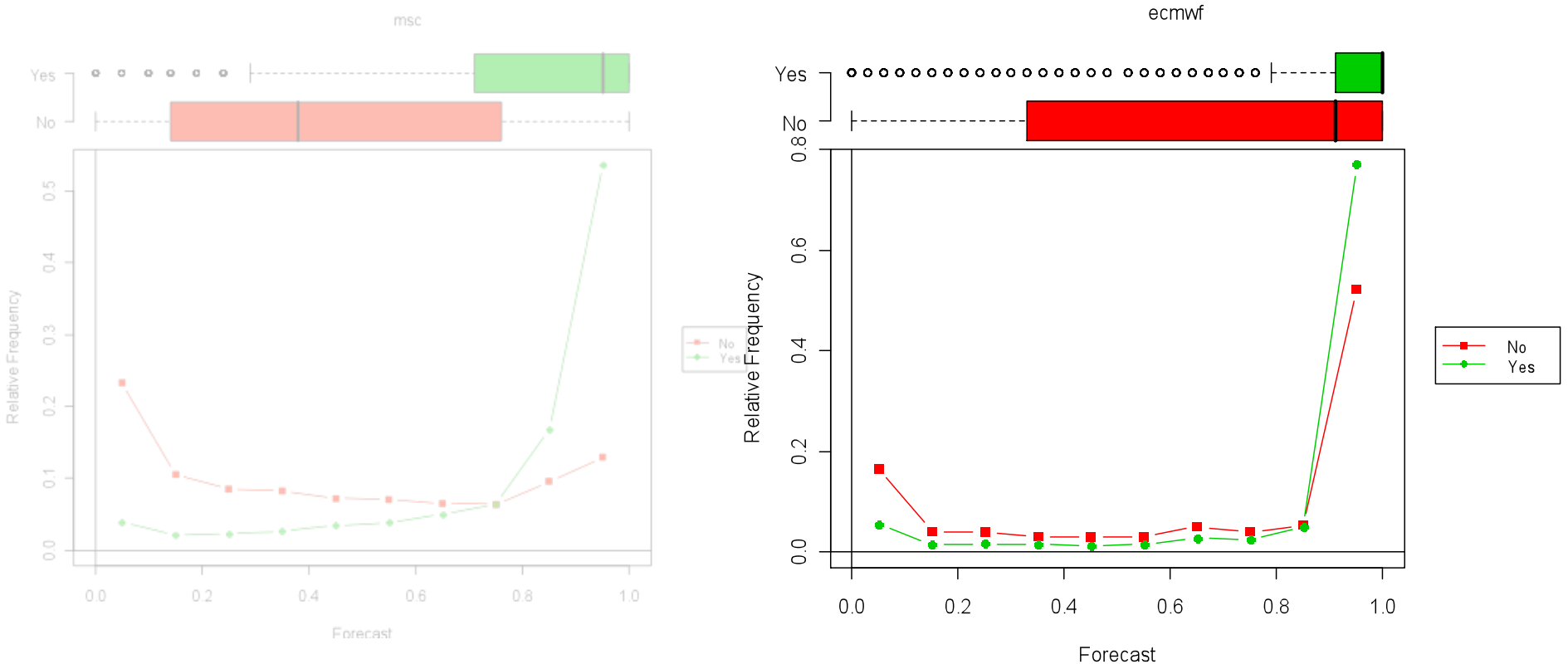
- *Discrimination*: The ability of the forecast system to clearly distinguish situations leading to the occurrence of an event of interest from those leading to the non-occurrence of the event.
- Depends on:
  - Separation of means of conditional distributions
  - Variance within conditional distributions



Good discrimination | Poor discrimination | Good discrimination

Sample Likelihood Diagrams: All precipitation, 20 Cdn stns, one year.

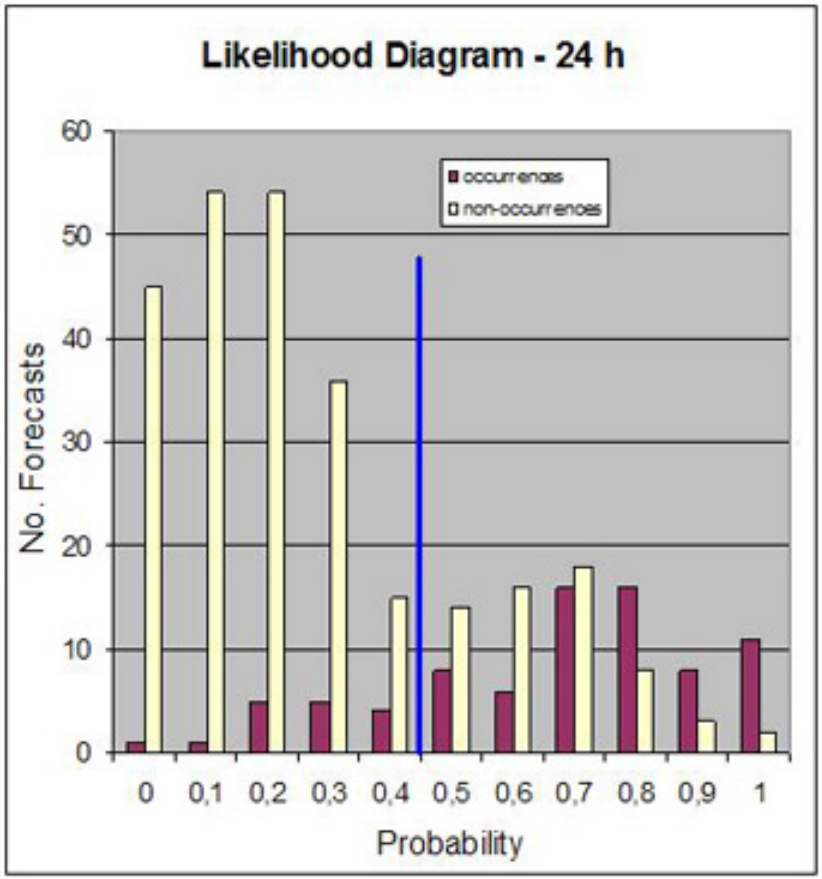
Discrimination: The ability of the forecast system to clearly distinguish situations leading to the occurrence of an event of interest from those leading to the non-occurrence of the event.

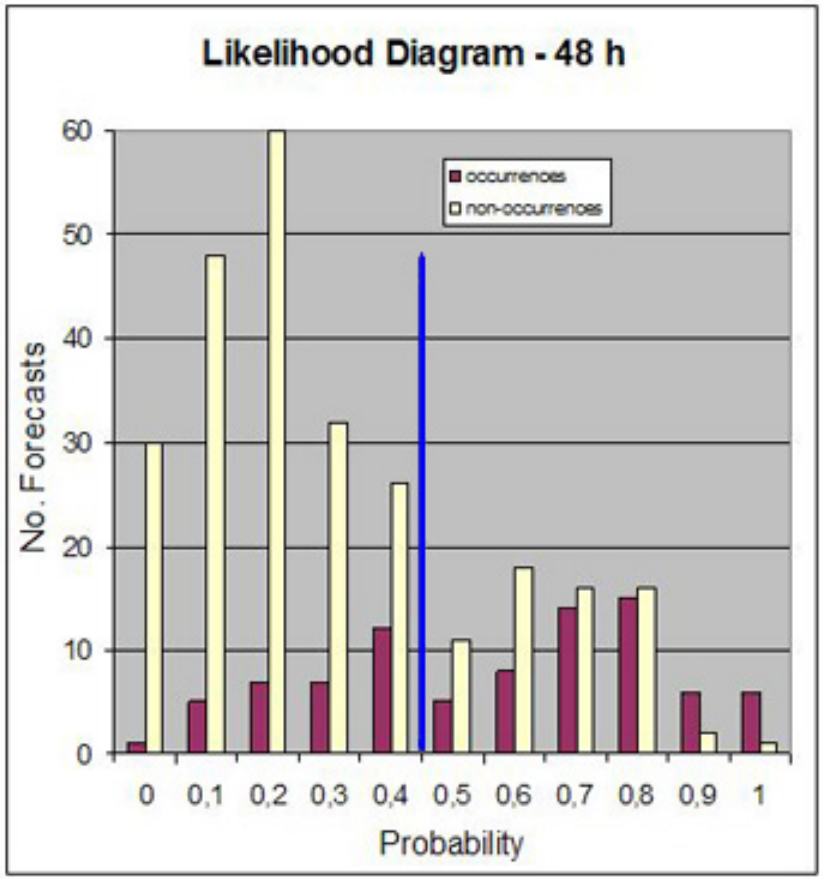


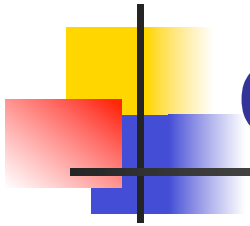
Relative Operating Characteristic curve: Construction

HR – Number of correct fcsts of event/total occurrences of event

FA – Number of false alarms/total occurrences of non-event



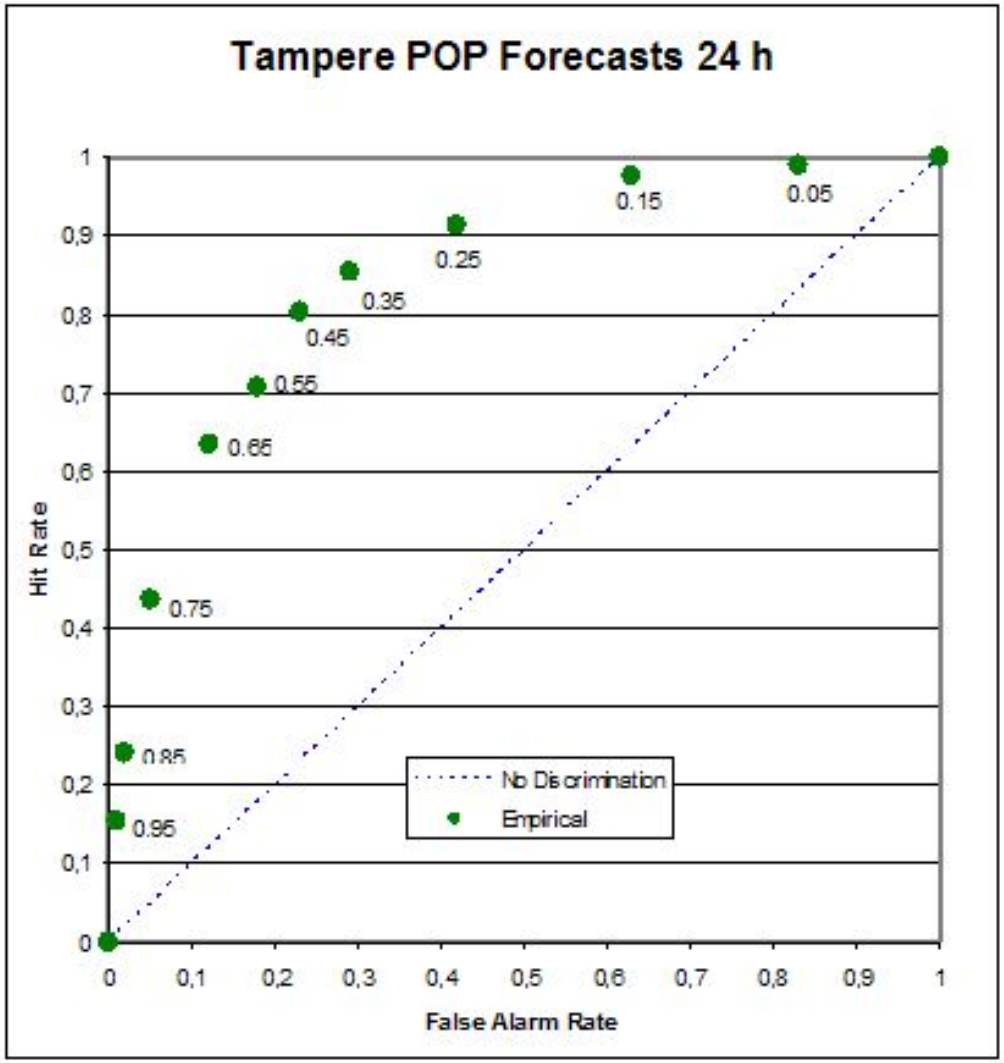




Construction of ROC curve

- From original dataset, determine bins
  - Can use binned data as for Reliability diagram BUT
  - There must be enough occurrences of the event to determine the conditional distribution given occurrences – may be difficult for rare events.
  - Generally need at least 5 bins.
- For each probability threshold, determine HR and FA
- Plot HR vs FA to give empirical ROC.
- Use binormal model to obtain ROC area; recommended whenever there is sufficient data >100 cases or so.
  - For small samples, recommended method is that described by Simon Mason. (See 2007 tutorial)

Empirical ROC



ROC - Interpretation

Interpretation of ROC:

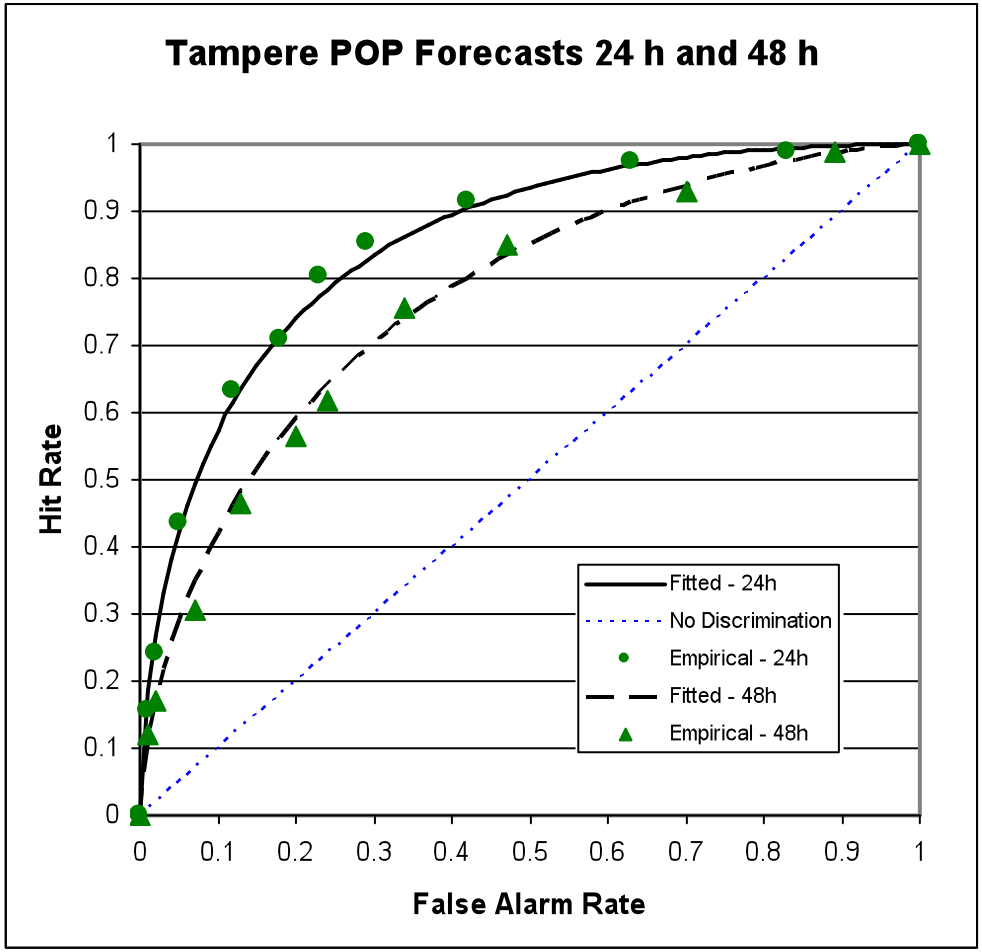
*Quantitative measure: Area under the curve – ROCA

*Positive if above 45 degree 'No discrimination' line where ROCA = 0.5

*Perfect is 1.0.

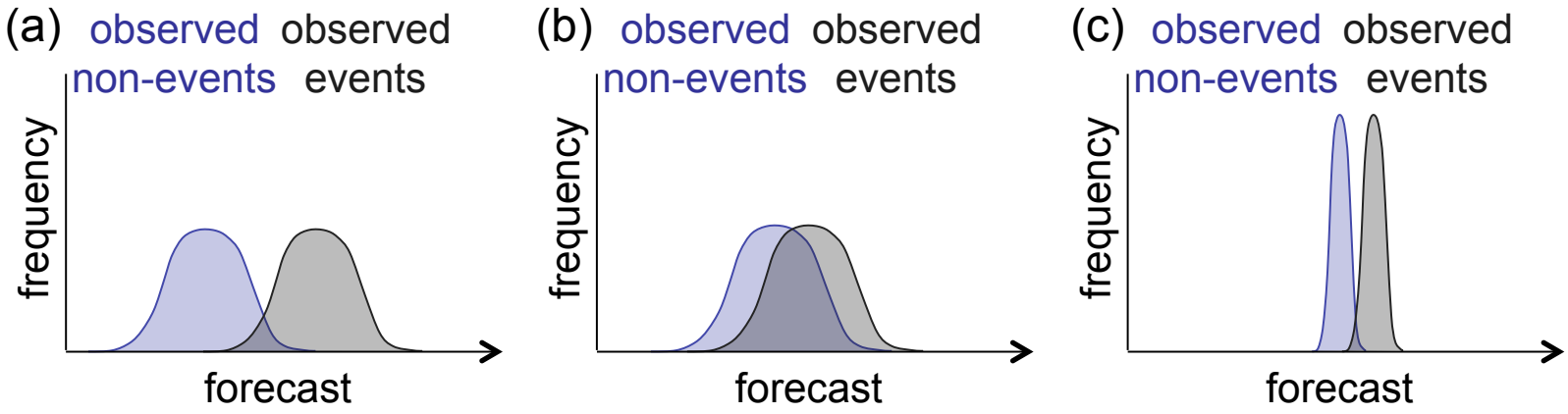
ROC is NOT sensitive to bias: It is necessarily only that the two conditional distributions are separate

* Can compare with deterministic forecast – one point



Discrimination

- Depends on:
  - Separation of means of conditional distributions
  - Variance within conditional distributions

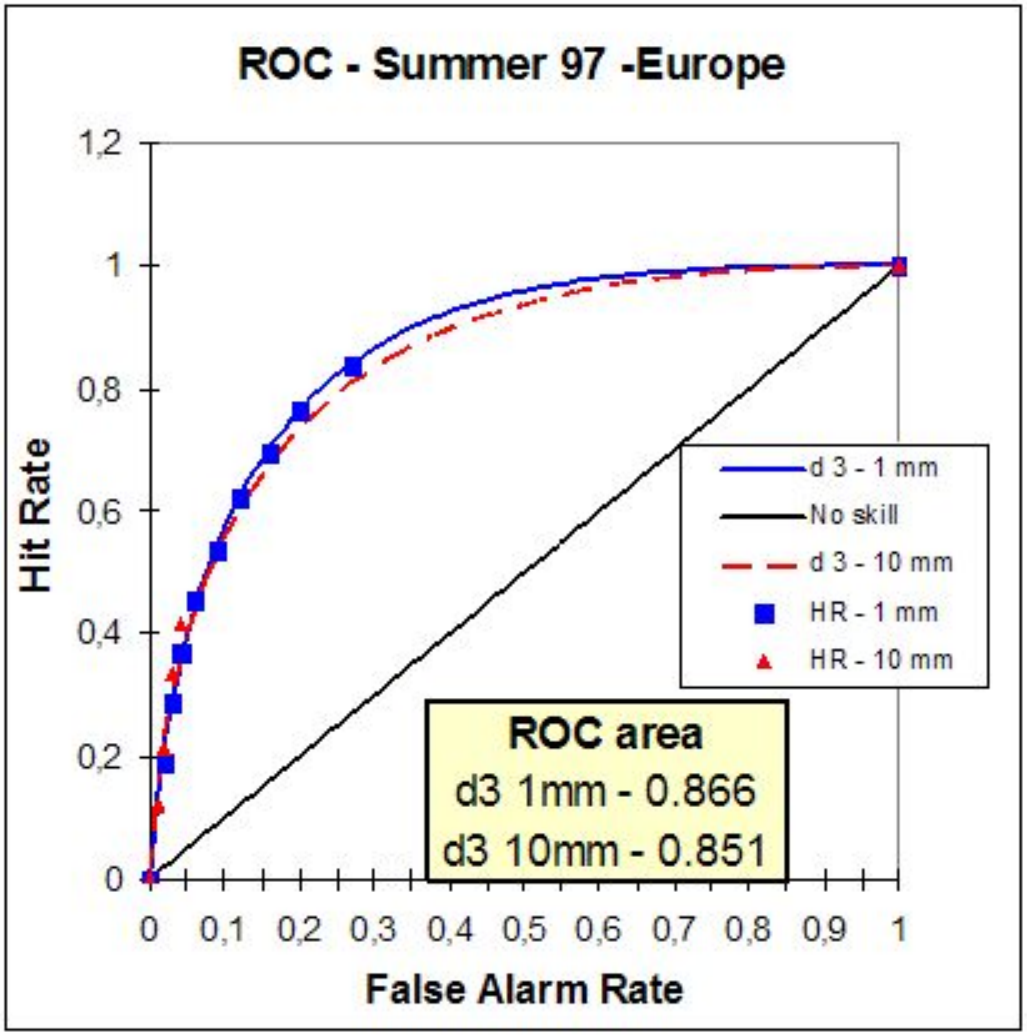


Good discrimination | Poor discrimination | Good discrimination

ROC for infrequent events

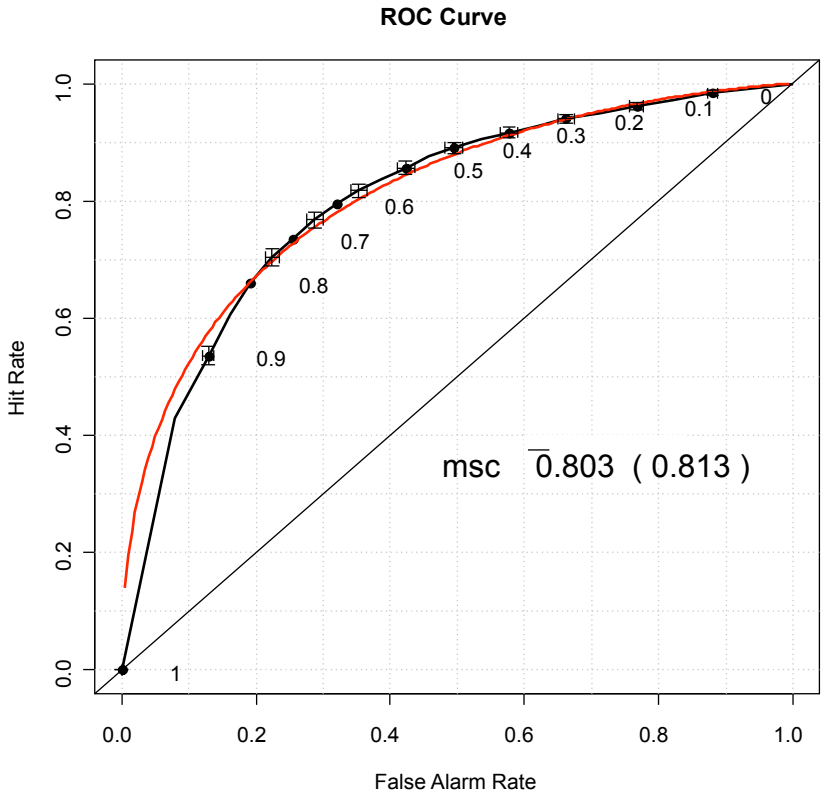
For fixed binning (e.g. deciles), points cluster towards lower left corner for rare events: subdivide lowest probability bin if possible.

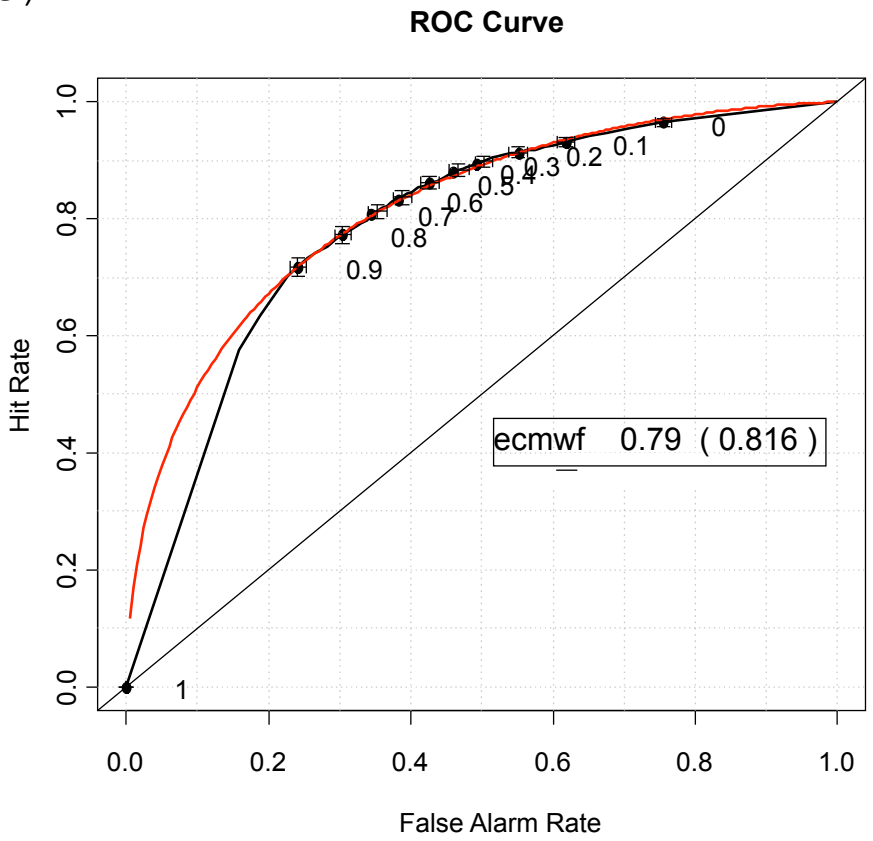
Remember that the ROC is insensitive to bias (calibration).

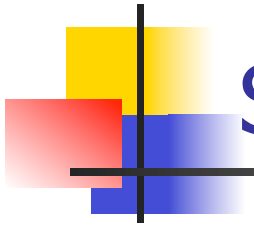


ROC in R

```
roc.plot.default(DAT$obs,  DAT$msc, binormal = TRUE,
legend = TRUE, leg.text = "msc", plot = "both", CI = TRUE)

roc.area(DAT$obs,  DAT$msc)
```







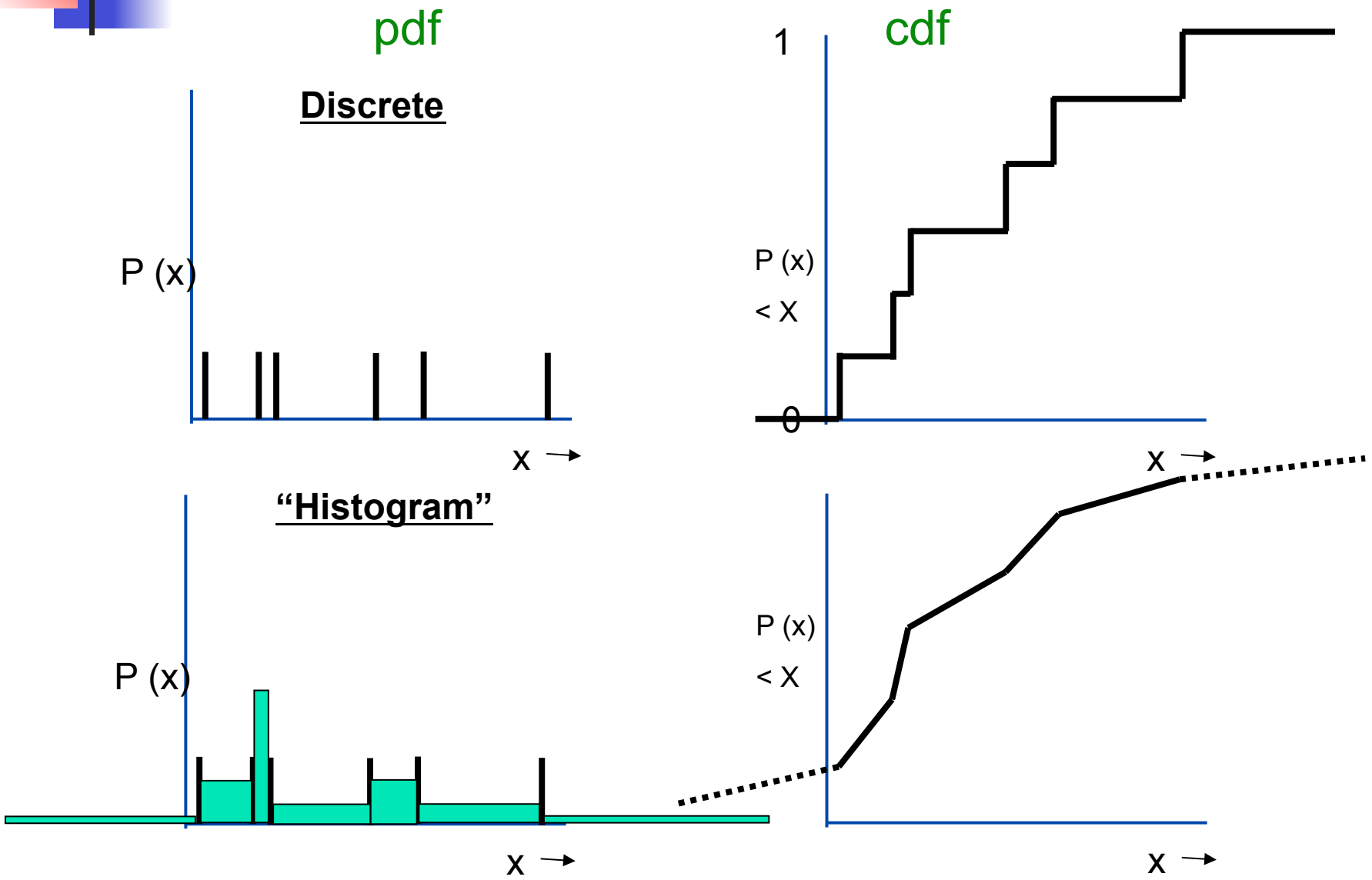
Summary - ROC

- Measures “discrimination”
- Plot of Hit rate vs false alarm rate
- Area under the curve – by fitted model
- Sensitive to sample climatology – careful about averaging over areas or time
- NOT sensitive to bias in probability forecasts – companion to reliability diagram
- Related to the assessment of “value” of forecasts
- Can compare directly the performance of probability and deterministic forecast

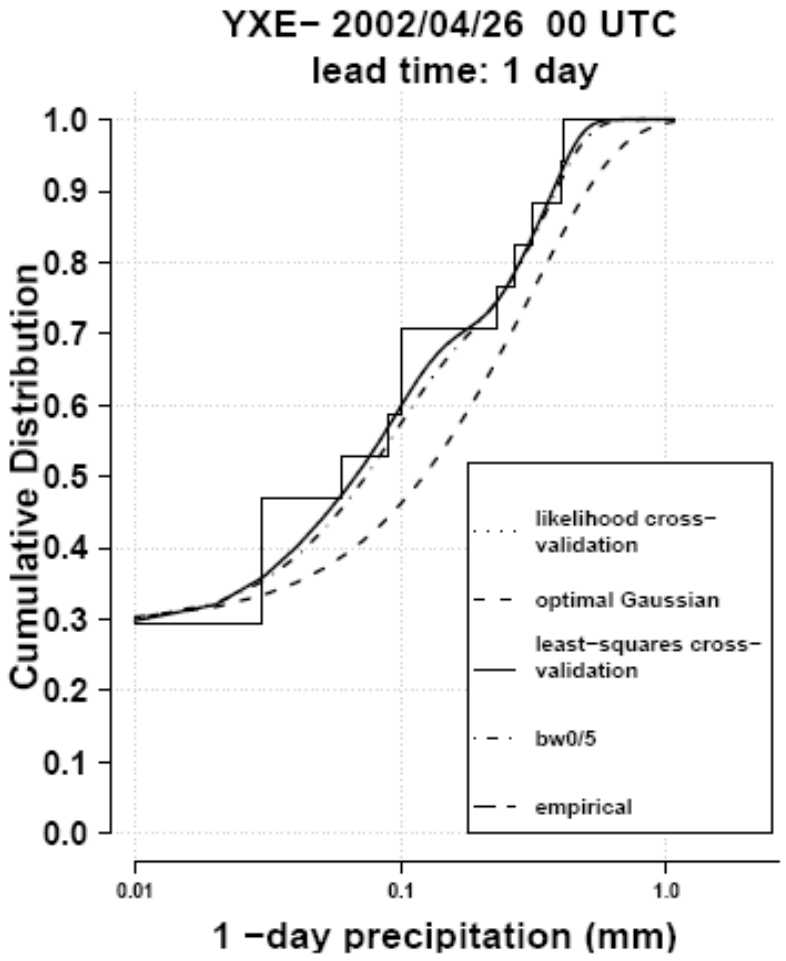
Data considerations for ensemble verification

- An extra dimension – many forecast values, one observation value
  - Suggests data matrix format needed; columns for the ensemble members and the observation, rows for each event
- Raw ensemble forecasts are a collection of deterministic forecasts
- The use of ensembles to generate probability forecasts requires interpretation.
  - i.e. processing of the raw ensemble data matrix.

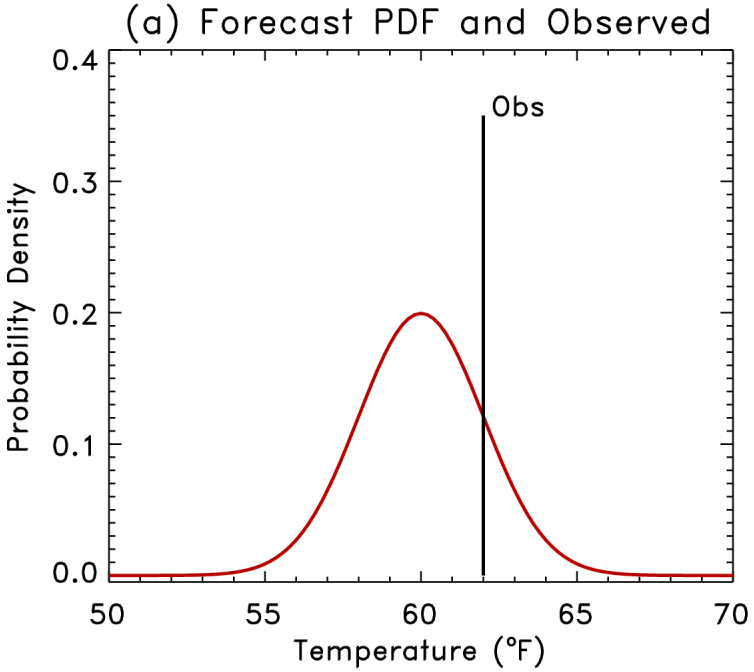
PDF interpretation from ensembles

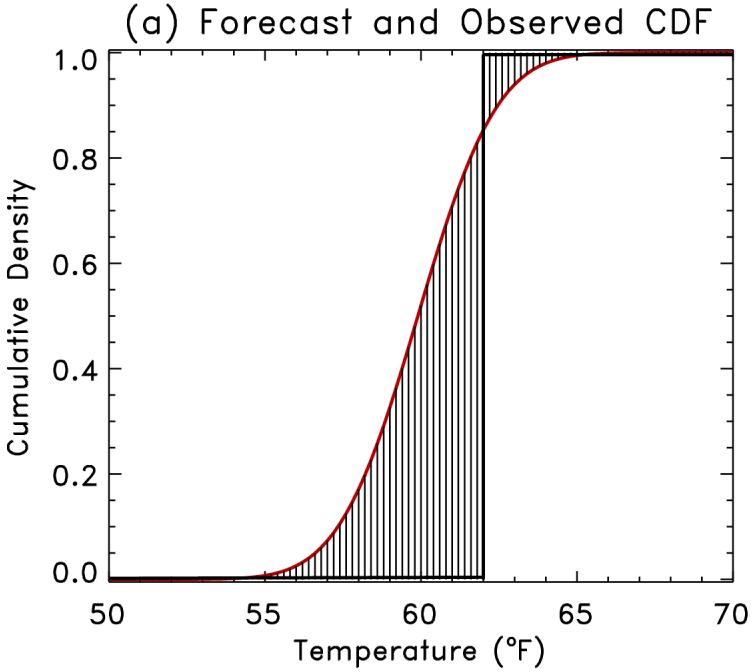


Example of discrete and fitted cdf

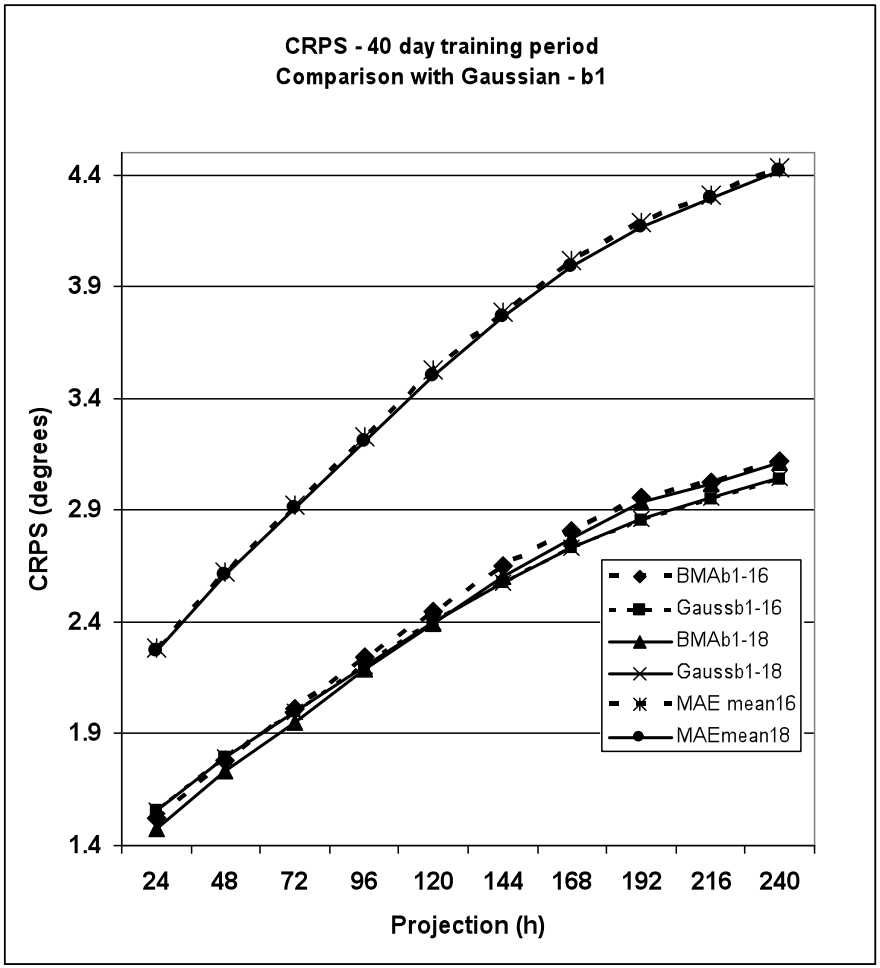


CRPS

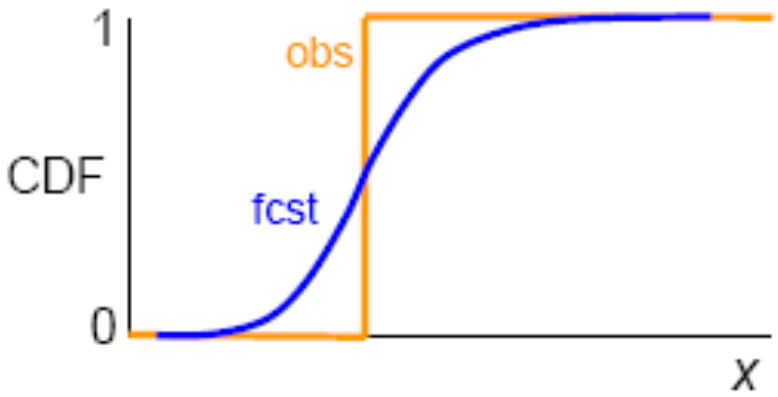

(a) Forecast PDF and Observed
Probability Density
0.0
0.1
0.2
0.3
0.4
Obs
50
55
60
65
70
Temperature (°F)


(a) Forecast and Observed CDF
Cumulative Density
0.0
0.2
0.4
0.6
0.8
1.0
50
55
60
65
70
Temperature (°F)

Continuous Rank Probability Score



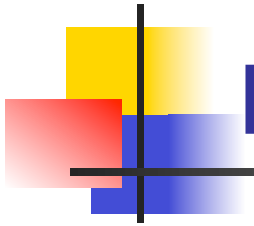
$$CRPS(P, x_a) = \int_{-\infty}^{\infty} \left[P(x) - P_a(x) \right]^2 dx$$



-difference between observation and forecast, expressed as cdfs

-defaults to MAE for deterministic fcst

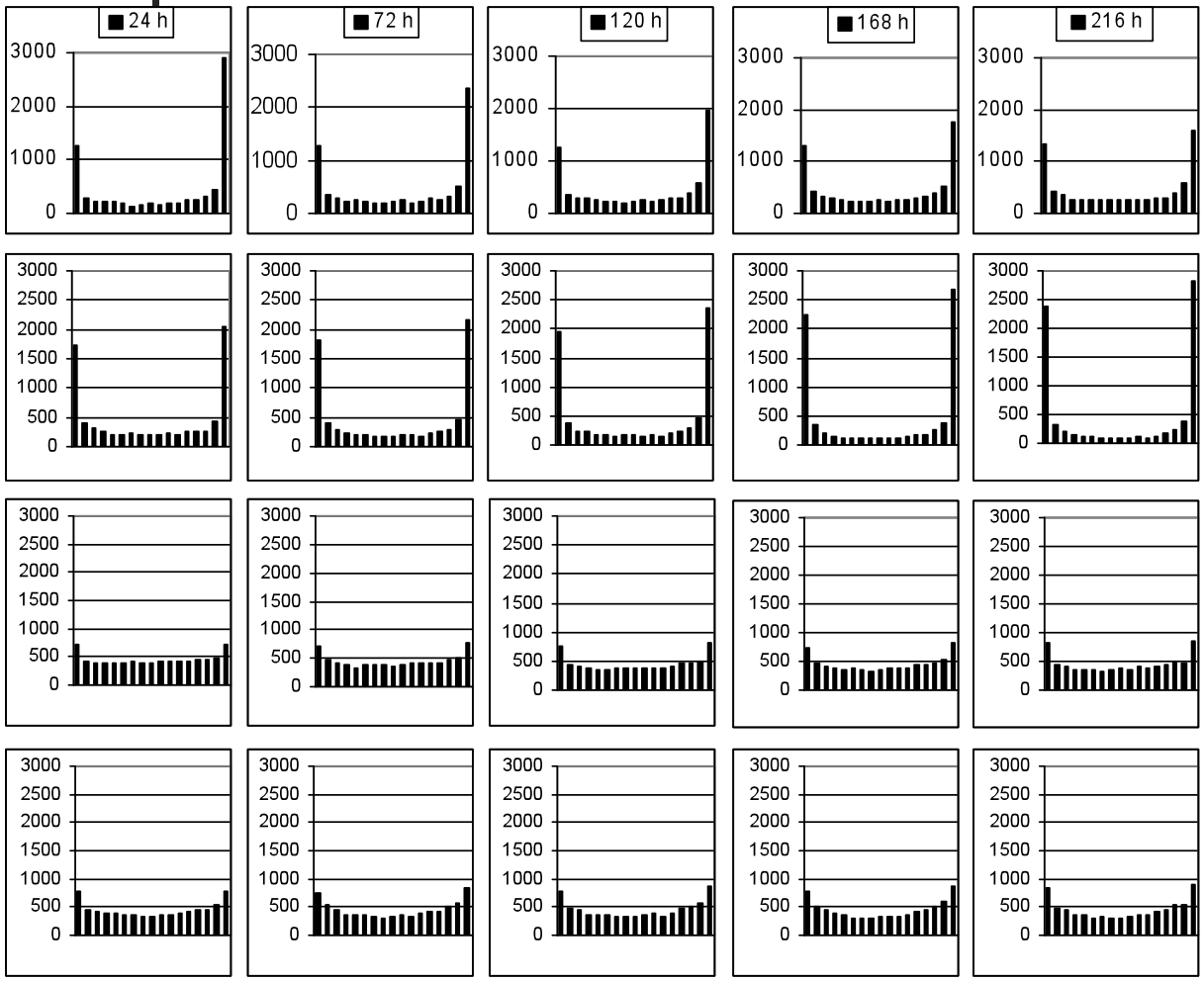
-flexible, can accommodate uncertain obs



Rank Histogram

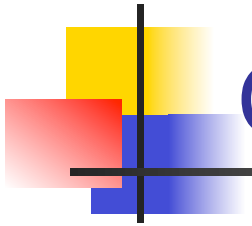
- Commonly used to diagnose the average spread of an ensemble compared to observations
- Computation: Identify rank of the observation compared to ranked ensemble forecasts
- Assumption: observation equally likely to occur in each of n+1 bins. (questionable?)
- Interpretation:

Quantification of "departure from flat"



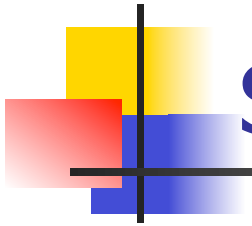
$$RMSD = \sqrt{\frac{1}{N+1}\sum_{k=1}^{N+1}\left(s_k - \frac{M}{N+1}\right)^2}$$

$$\sqrt{\frac{MN}{(N+1)^2}}$$



Comments on Rank Histogram

- Can quantify the “departure from flat”
- Not a “real” verification measure
- Who are the users?



Summary

- Summary score: Brier and Brier Skill
  - Partition of the Brier score
- Reliability diagrams: Reliability, resolution and sharpness
- ROC: Discrimination
- Diagnostic verification: Reliability and ROC
- Ensemble forecasts: Summary score - CRPS