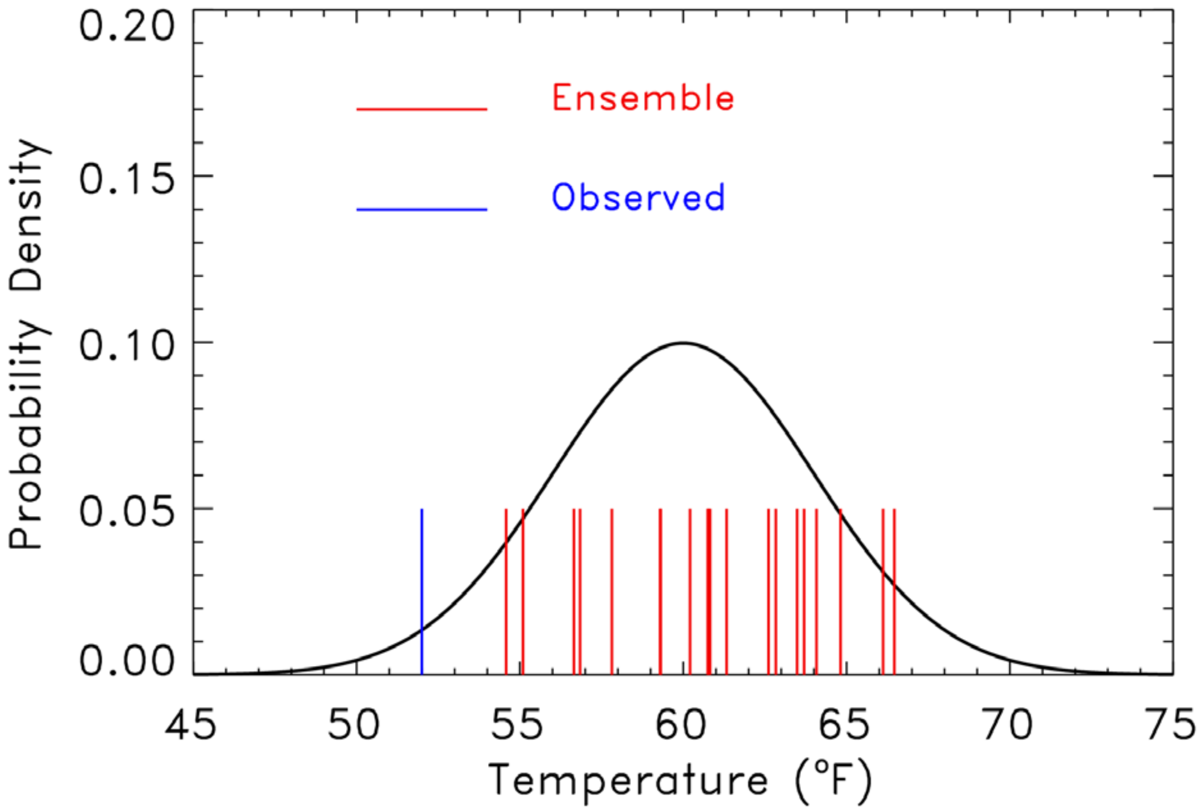
Verification of ensembles

Barbara G. Brown

Acknowledgments: Tom Hamill, Laurence Wilson, Tressa Fowler

How good is this ensemble forecast?



Questions to ask before beginning?

- How were the ensembles constructed?
  - Poor man's ensemble (distinct members)
  - Multi-physics (distinct members)
  - Random perturbation of initial conditions (anonymous members)
- How are your forecasts used?
  - Improved point forecast (ensemble mean)
  - Probability of an event
  - Full distribution

Approaches to evaluating ensemble forecasts

- As individual members
  - Use methods for continuous or categorical forecasts
- As probability forecasts
  - Create probabilities by applying thresholds or statistical post-processing
- As a full distribution
  - Use individual members or fit a distributions through post-processing

Evaluate each member as a separate, deterministic forecast

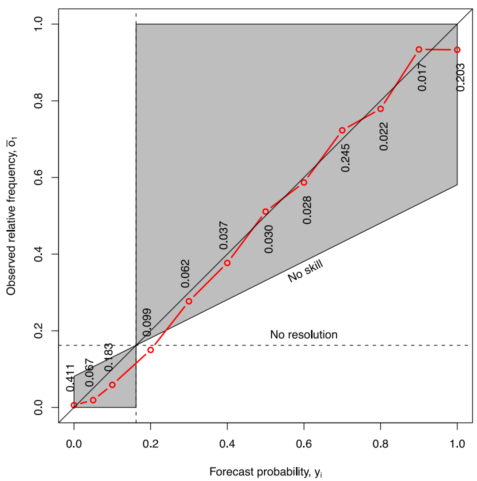
- Why? Because it is easy and important
    - If members are unique, it might provide useful diagnostics.
    - If members are biased, verification statistics might be skewed.
    - If members have different levels of bias, should you calibrate?
  - Do these results conform to your understanding of how the ensemble members were created?

Verifying a probabilistic forecast

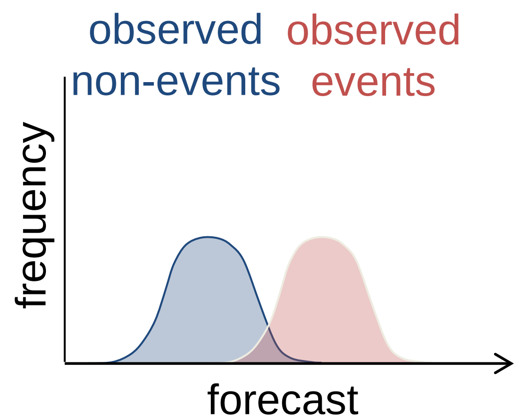
- You cannot verify a probabilistic forecast with a single observation.
- The more data you have for verification, (as with other statistics) the more certain you are.
- Rare events (low probability) require more data to verify.
- These comments refer to probabilistic forecasts developed by methods other than ensembles as well.

Properties of a perfect probabilistic forecast of a binary event.

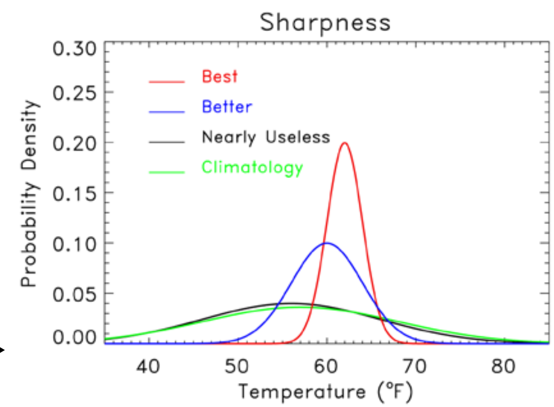
Reliability



Resolution



Sharpness



The Brier Score

- Mean square error of a probability forecast

$$BS = \frac{1}{n}\sum_{i=1}^{n}(f_i - x_i)^2$$

where n is the number of forecasts

f_i is the forecast prob on occasion i

x_i is the observation (0 or 1) on occasion i

- Weights larger errors more than smaller ones

Brier Score

$$\text{BS} = \frac{1}{n}\sum_{k=1}^{n}(f_k - x_k)^2$$ where

f_k = forecast probability on occasion k

x_k = observation (0 or 1) on occasion k

BS can be decomposed into 3 components that represent important properties of the forecasts:

$$\text{BS} = \underbrace{\frac{1}{n}\sum_{i=1}^{I} N_i (f_i - \overline{x}_i)^2}_{\text{Reliability}} - \underbrace{\frac{1}{n}\sum_{i=1}^{I} N_i (\overline{x}_i - \overline{x})^2}_{\text{Resolution}} + \underbrace{\overline{x}(1-\overline{x})}_{\text{Uncertainty}}$$

Where the I is the number of discrete values of f (e.g., f_1 = 0.05, f_2 = 0.10, f_3 = 0.20, ... etc.) and

$$n = \sum_{i=1}^{I} N_i \qquad \overline{x}_i = \frac{1}{N_i}\sum_{k \in N_i} x_k$$ and $$\overline{x} = \frac{1}{n}\sum_{k=1}^{n} x_k = \frac{1}{N}\sum_{i=1}^{I} N_i \overline{x}_i$$

Components of the Brier Score

- Reliability

  Measures how well the conditional relative frequency of events matches the forecast

$$\frac{1}{n}\sum_{i=1}^{I} N_i (f_i - \overline{x}_i)^2$$

- Resolution

  Measures how well the forecasts distinguish situations with different frequencies of occurrence

$$\frac{1}{n}\sum_{i=1}^{I} N_i (\overline{x}_i - \overline{x})^2$$

- Uncertainty

  Measures the variability in the observations (i.e., the difficulty of the forecast situations)

$$\overline{x}(1-\overline{x})$$

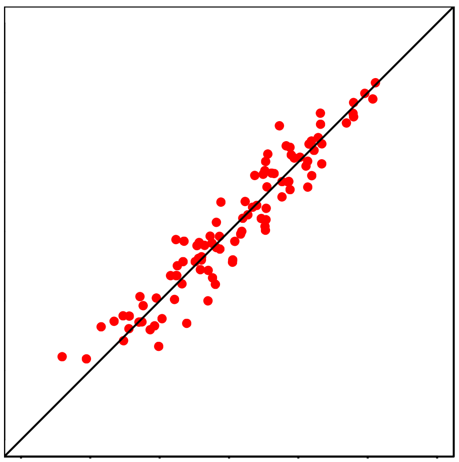
Looking at Brier Score *components* is critical to understand forecast performance

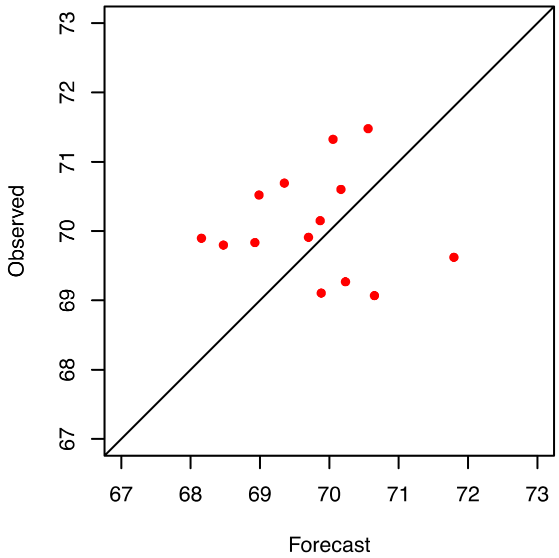
Brier Skill Score (BSS)

$$\mathrm{BSS} = \frac{\mathrm{RES} - \mathrm{REL}}{\mathrm{UNC}}$$

BSS is a simple combination of the 3 components of the Brier Score (assumes "Sample Climatology" as the reference forecast)

Our friend, the scatterplot

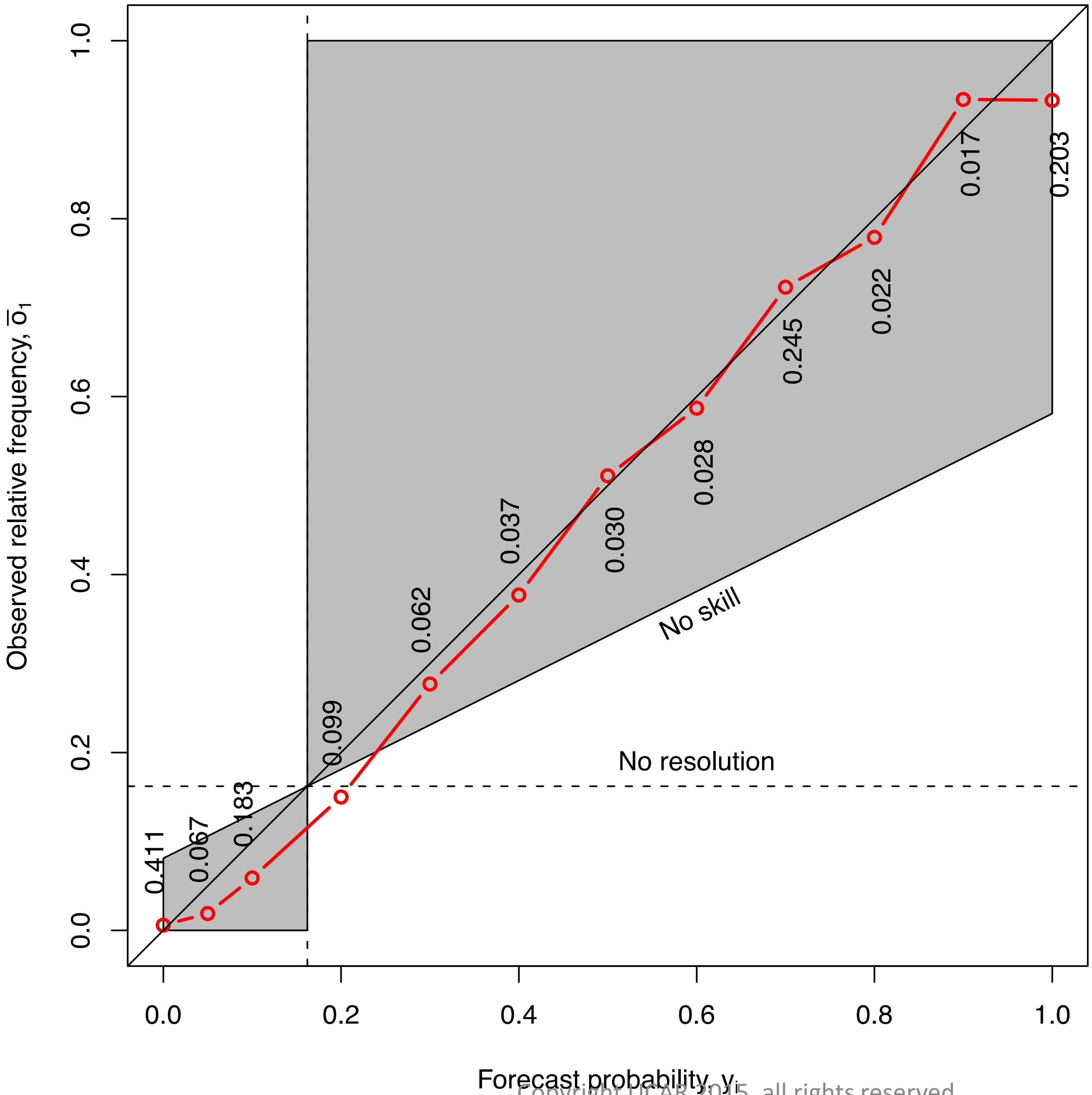




Introducing the attribute diagram!

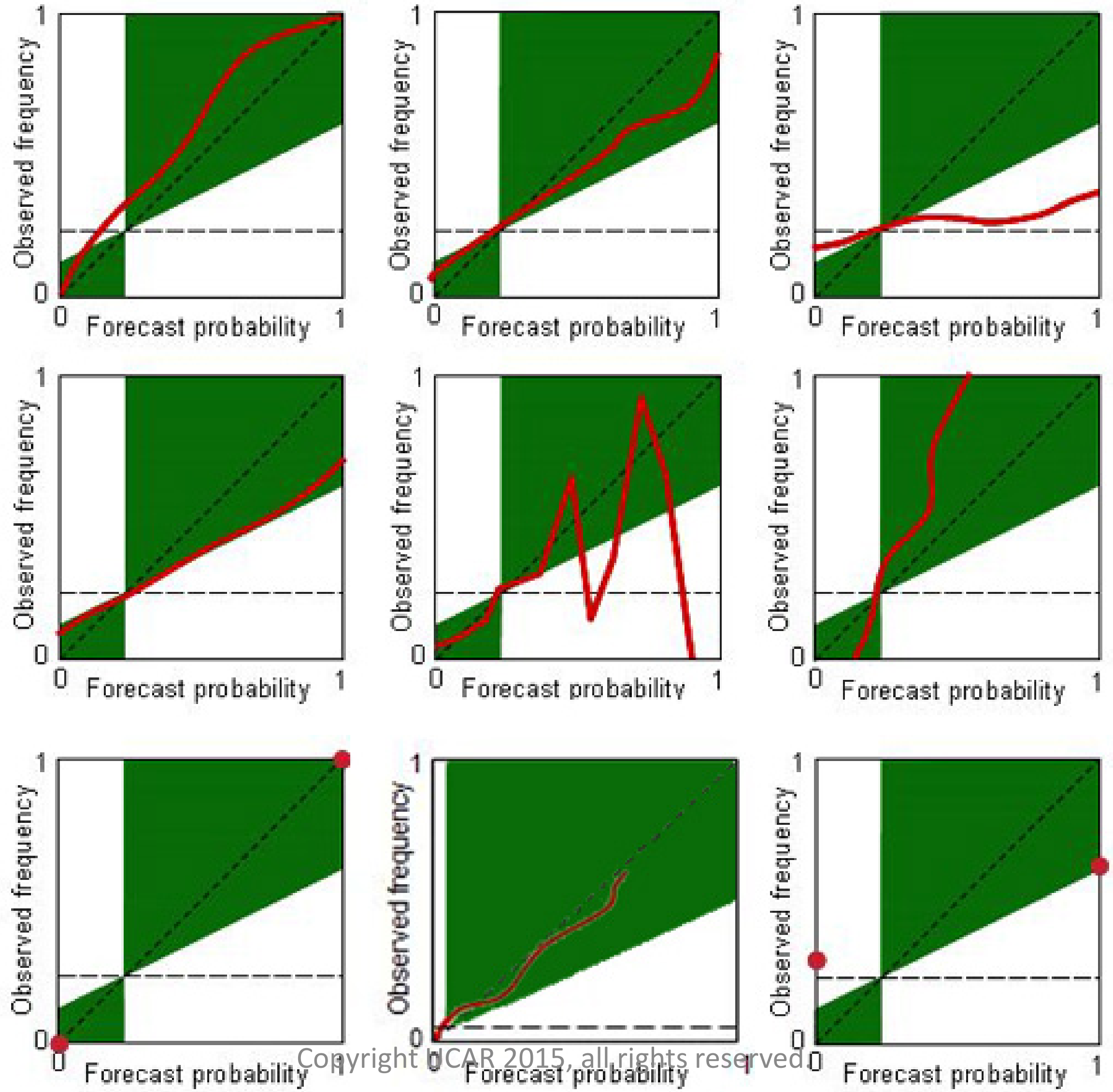
(close relative to the reliability diagram)

- Analogous to the scatter plot- same intuition holds.
- Data must be binned!
- Hides how much data is represented by each
- Expresses conditional probabilities.
- Confidence intervals can illustrate the problems with small sample sizes.



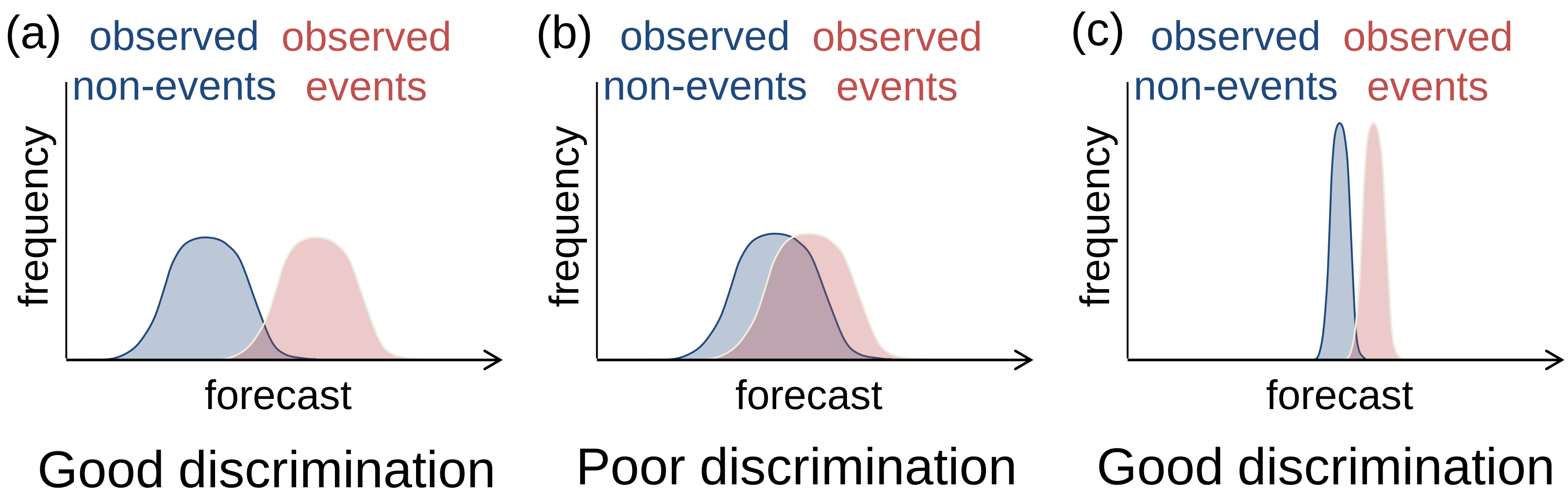
Attribute diagram shows reliability, resolution, skill

Reliability Diagram Exercise



Discrimination

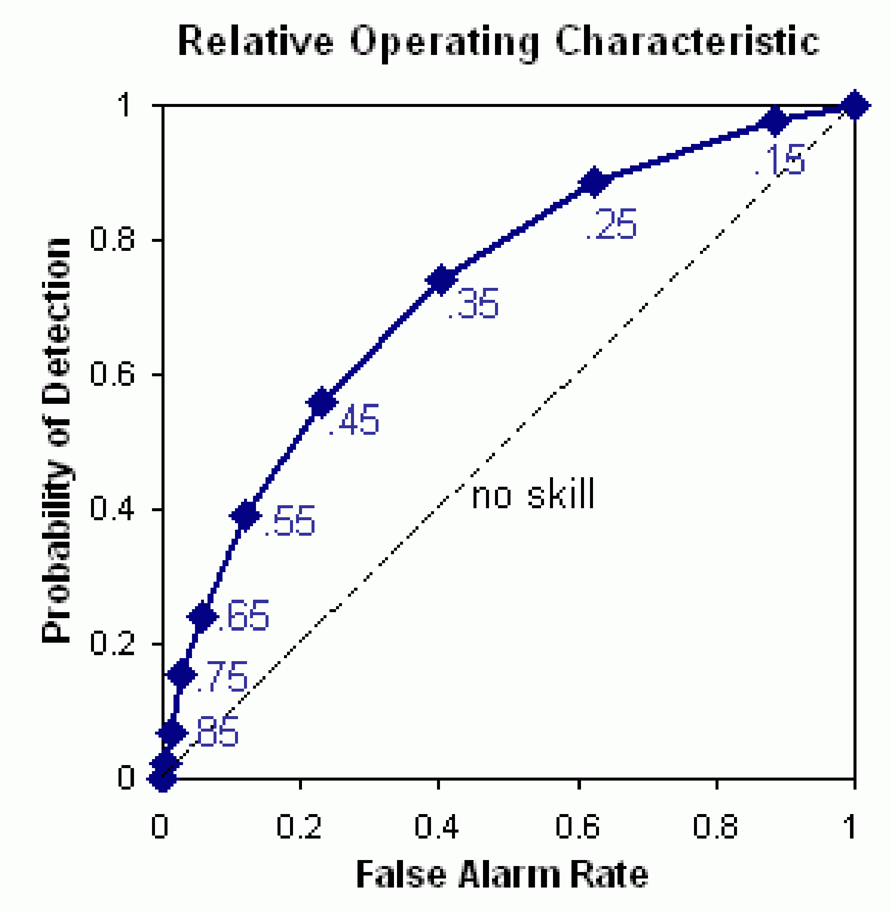
- *Discrimination*: The ability of the forecast system to clearly distinguish situations leading to the occurrence of an event of interest from those leading to the non-occurrence of the event.
- Depends on:
  - Separation of means of conditional distributions
  - Variance within conditional distributions



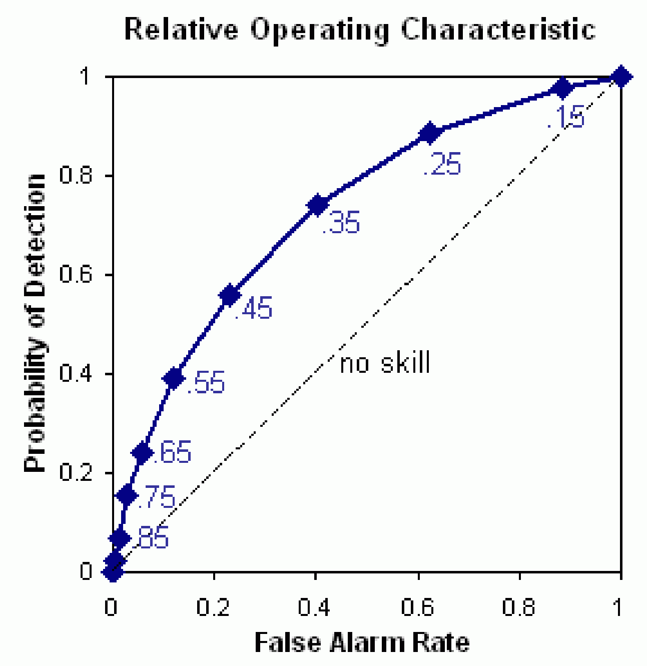
Relative Operating Characteristic (ROC)

Measures the ability of the forecast to discriminate between events and non-events (resolution)

→ *Plot hit rate H vs false alarm rate F using a set of varying probability thresholds to make the yes/no decision*.



Interpretation of ROC

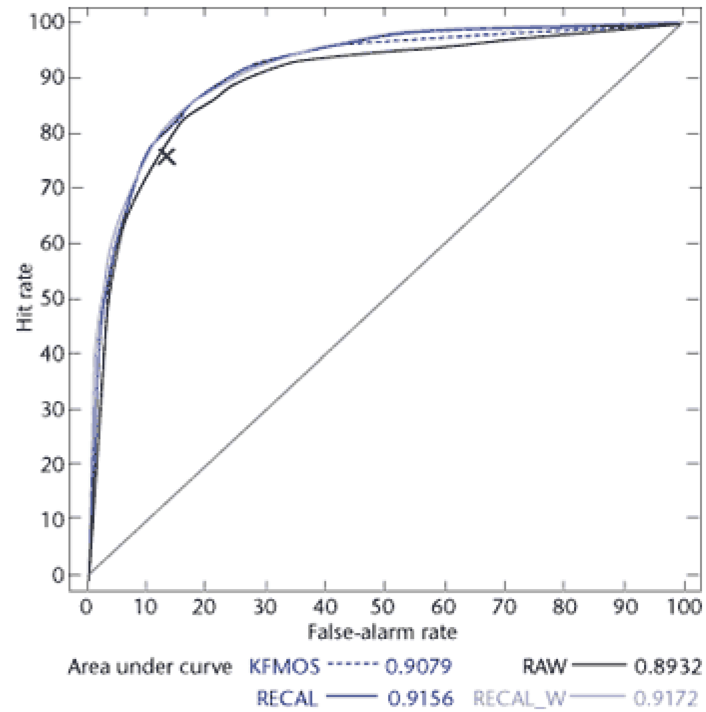


- Close to upper left corner – *good resolution*
- Close to diagonal – *little skill*
- **Area under curve** ("ROC area") is a useful summary measure of forecast skill
- **Perfect**: ROC area = 1
- **No skill**: ROC area = 0.5
- ROC skill score ROCS = 2(ROCarea-0.5)
- *Not sensitive to bias.*

- ROC is **conditioned on the observations** (i.e., given that Y occurred, what was the corresponding forecast?)
- Reliability and ROC diagrams are good companions

Relative Operating Characteristic (ROC)

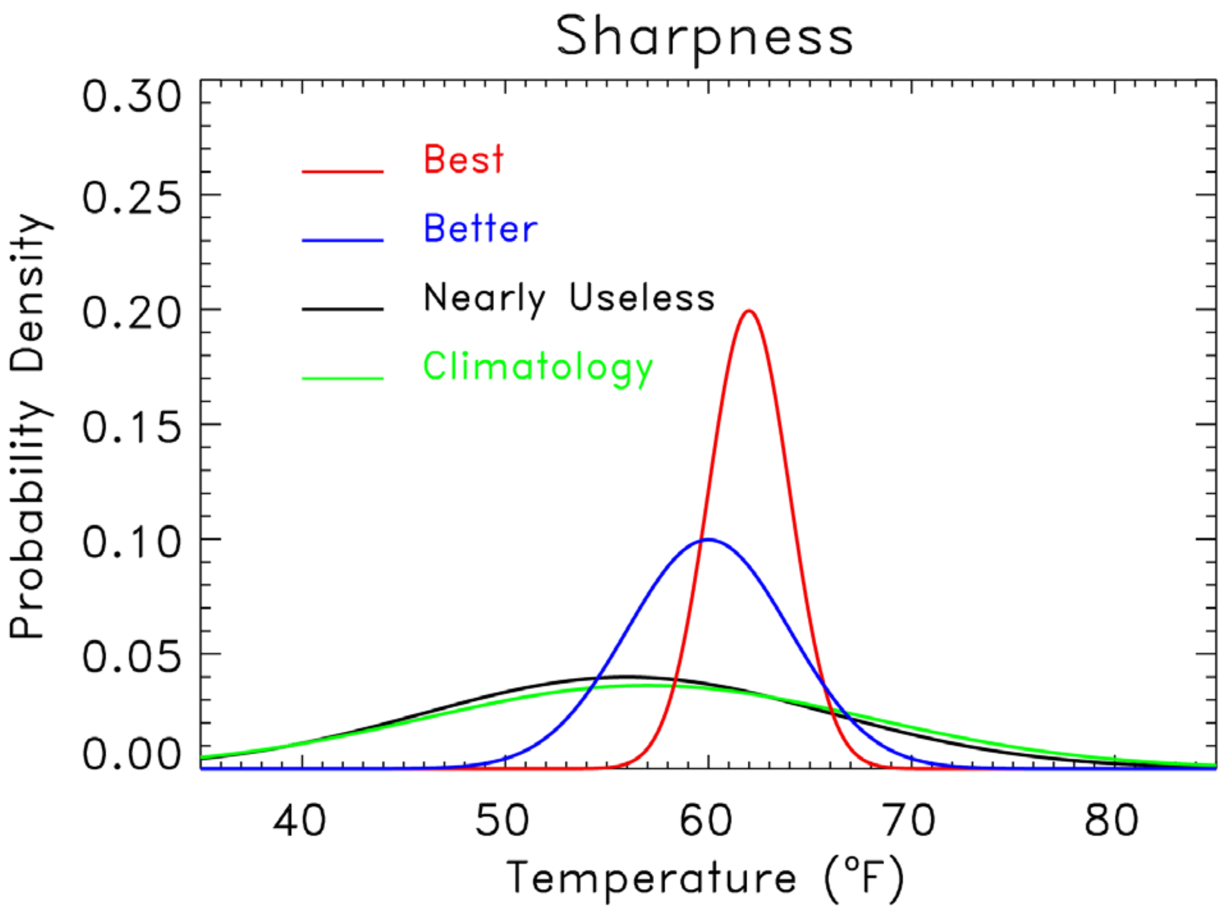
ROC example:



ROC diagram for T12< 5 °C at T+72. Shades indicate the different levels of statistical processing applied as shown in the key. The cross indicates the ROC (FAR, HR) of the ECMWF high-resolution deterministic model.

from "Verification of PREVIN site-specific probability forecasts", Met Office (http://www.metoffice.com/research/nwp/publications/nwp_gazette/dec01/verif.html)

Sharpness also important



"**Sharpness**" measures the specificity of the probabilistic forecast. Given two reliable forecast systems, the one producing the sharper forecasts is preferable.

But: don't want sharp if not reliable. Implies unrealistic confidence.

Sharpness ≠ resolution

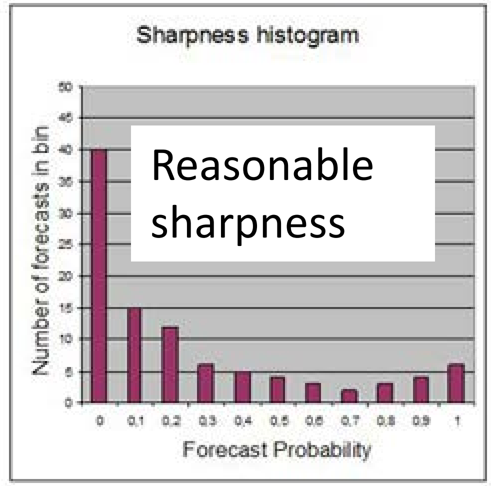
- Sharpness is *a property of the forecasts alone*; a measure of sharpness in Brier score decomposition would be how populated the extreme N_i*'s* are.

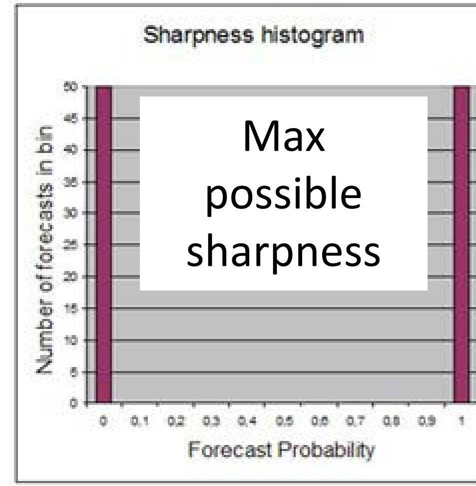
$$\mathrm{BS} = \frac{1}{n}\sum_{i=1}^{I} N_i (f_i - \overline{x}_i)^2 - \frac{1}{n}\sum_{i=1}^{I} N_i (\overline{x}_i - \overline{x})^2 + \overline{x}(1-\overline{x})$$

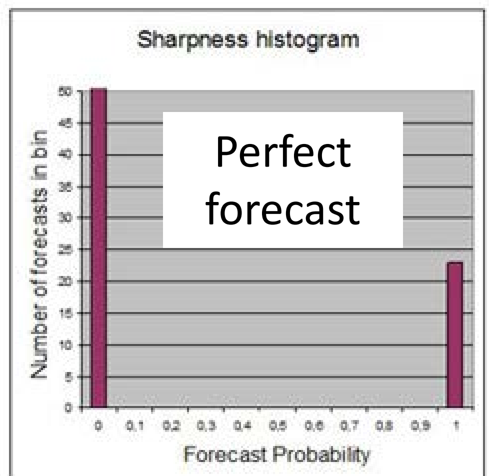
Sharpness for binary probability forecasts

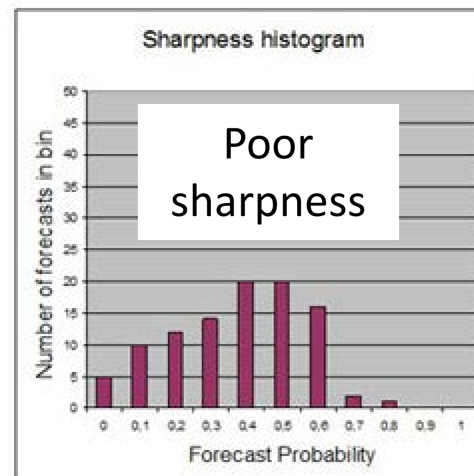
For a binary probability forecast, sharpness is based on the distribution (histogram) of frequencies associated with each possible probability

Sometimes summarized using the variance of the distribution







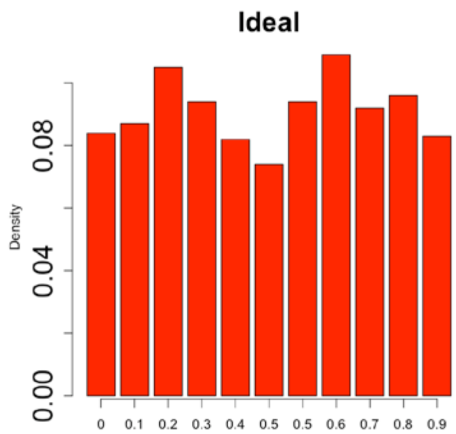


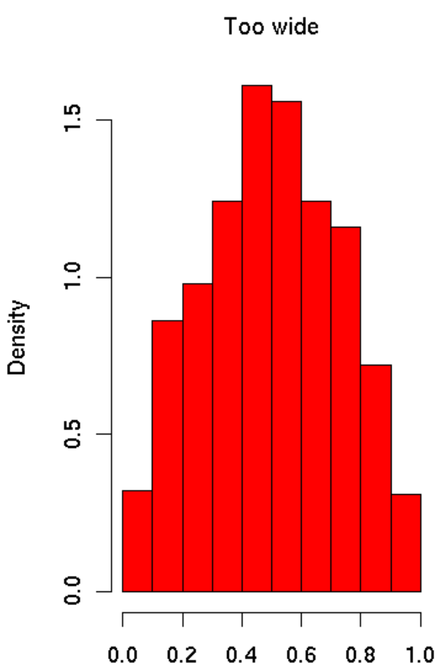
Forecasts of a full distribution

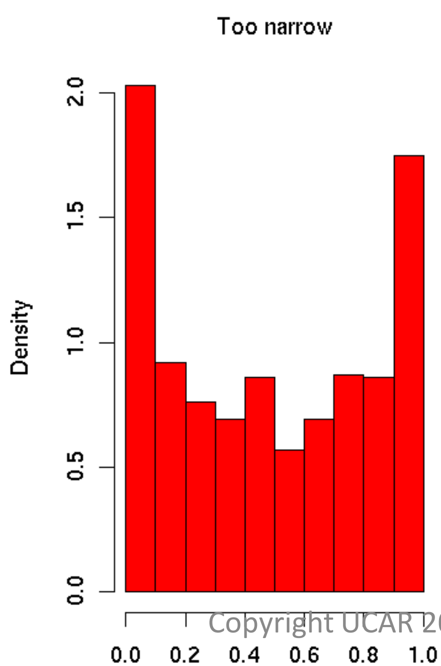
- How is it expressed?
  - Discretely by providing forecasts from all ensemble members
  - A parametric distribution – normal (ensemble mean, spread)
  - Smoothed function – kernel smoother

Evaluating ensembles

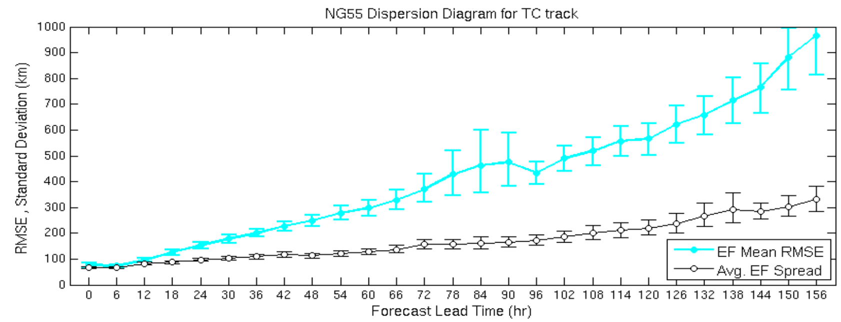
Rank Histo-
grams

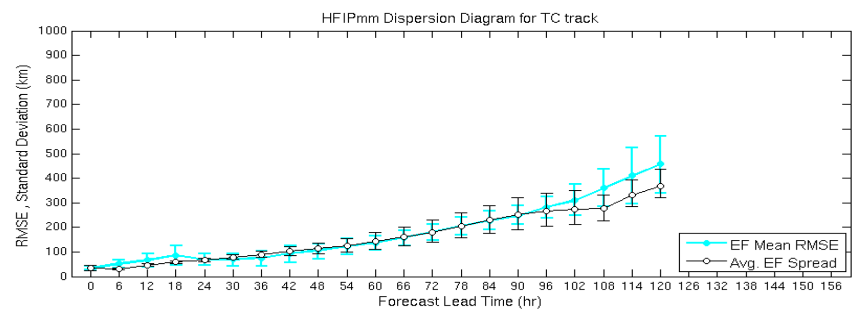






Spread-skill



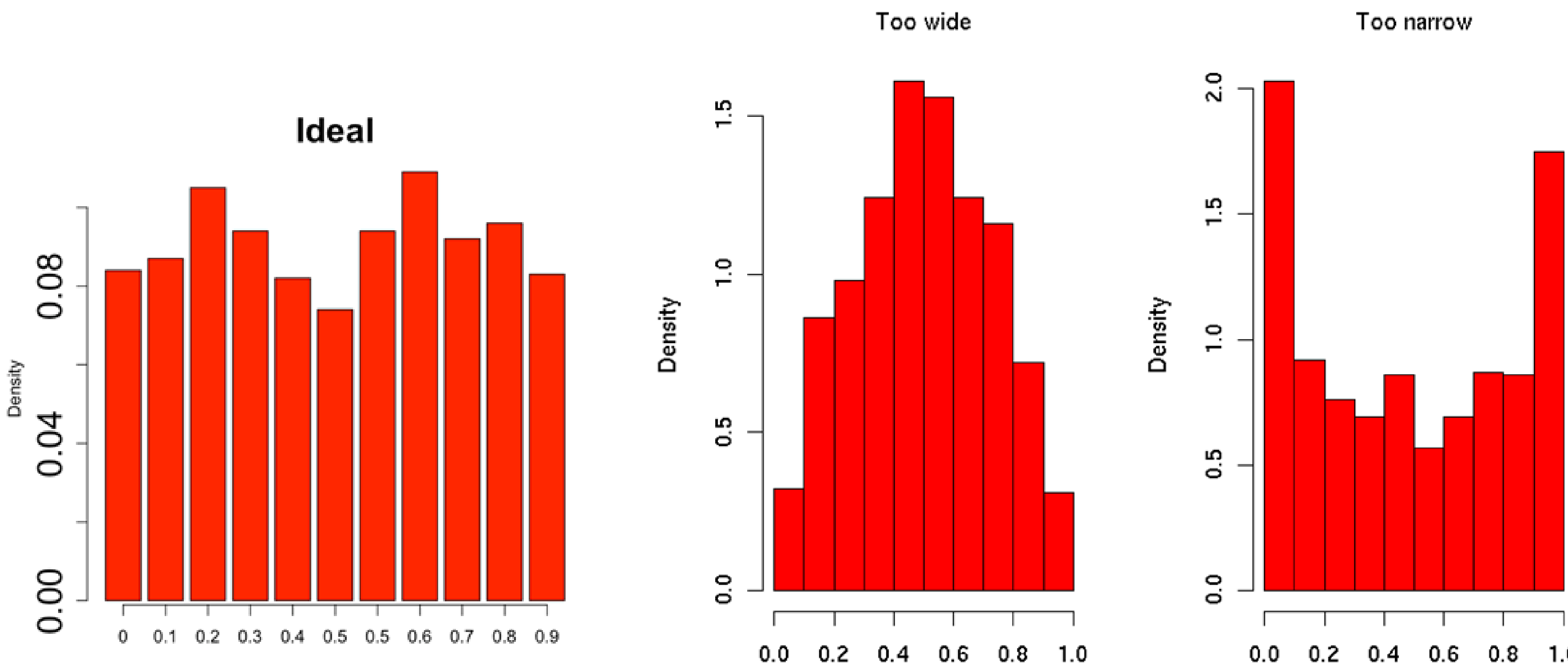


Continuous Ranked Probability Score: Measures skill using squared error (analogous to MAE)

Ensemble Calibration / Reliability

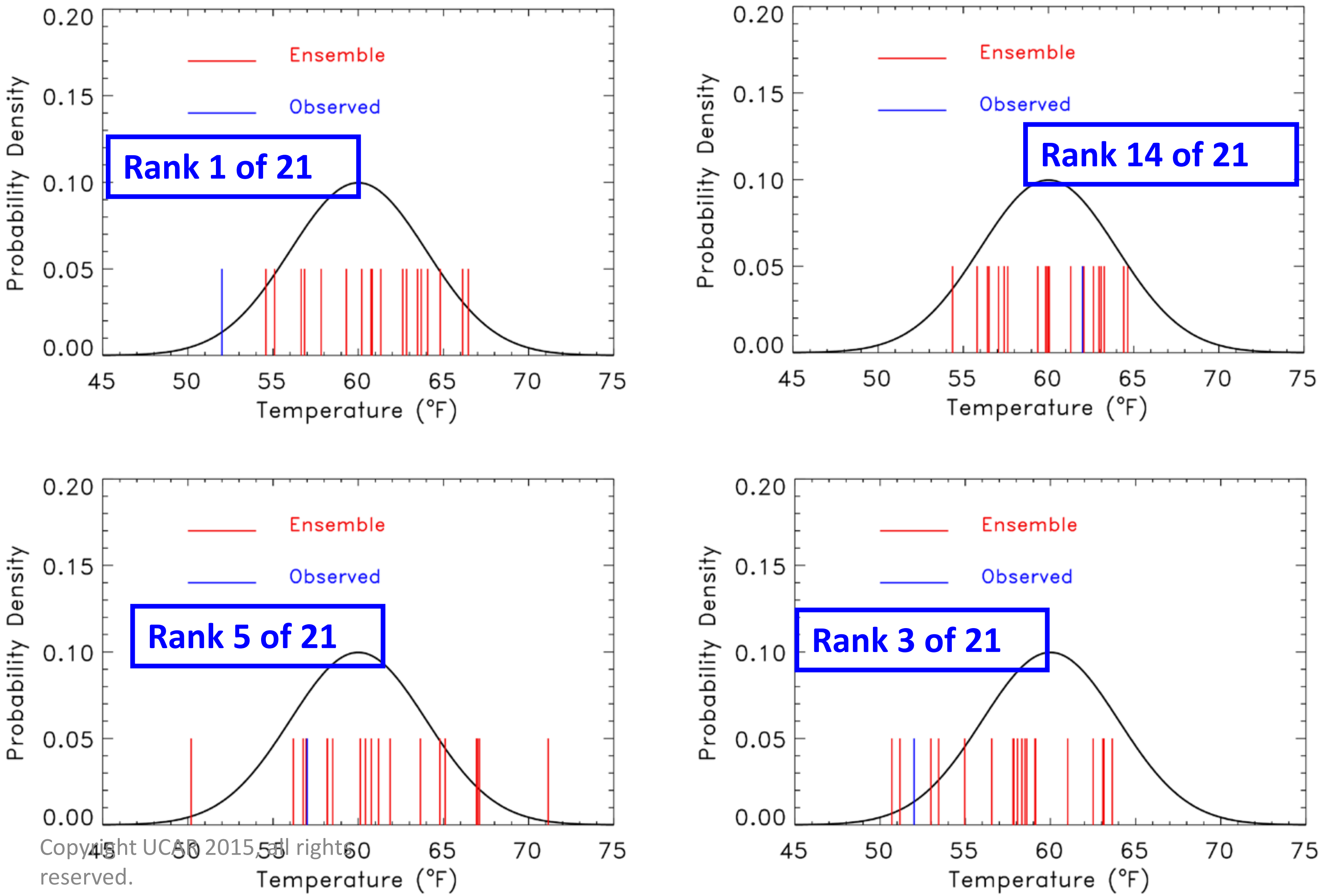
- By default, we assume all ensemble forecasts have the same number of members. Comparing forecasts with different number of members is an advanced topic.
- For a perfect ensemble, the observation comes from the same distribution as the ensemble.

Rank histogram examples

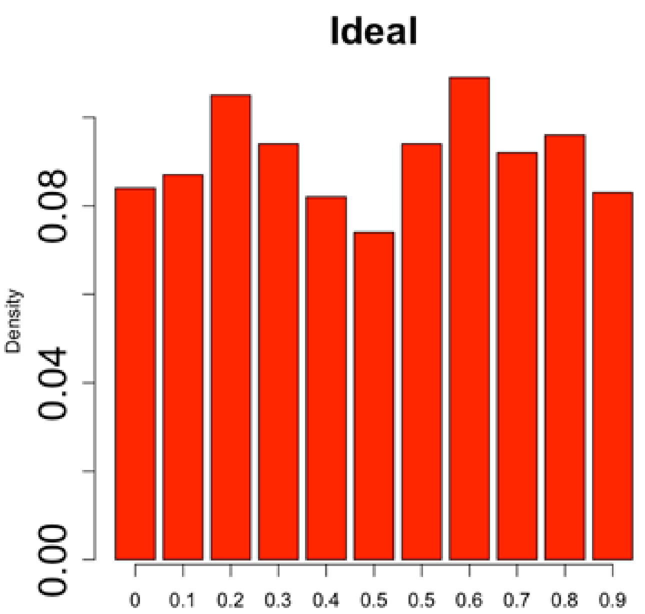


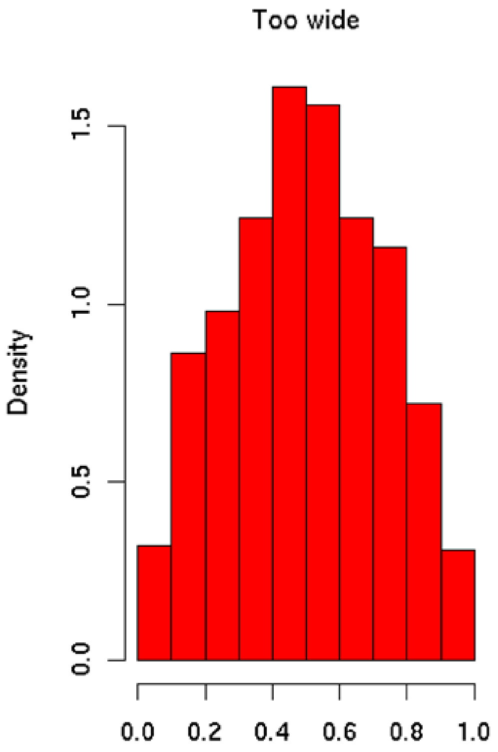
Rank histograms are a way to examine the calibration of an ensemble

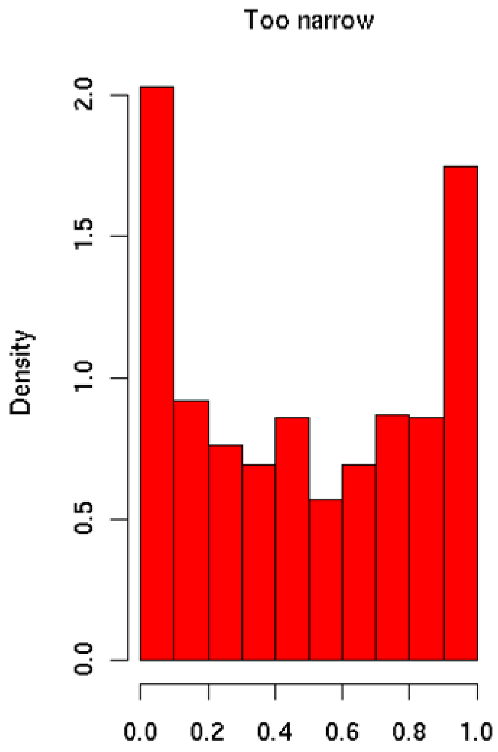
Creating rank histograms



Rank histogram examples





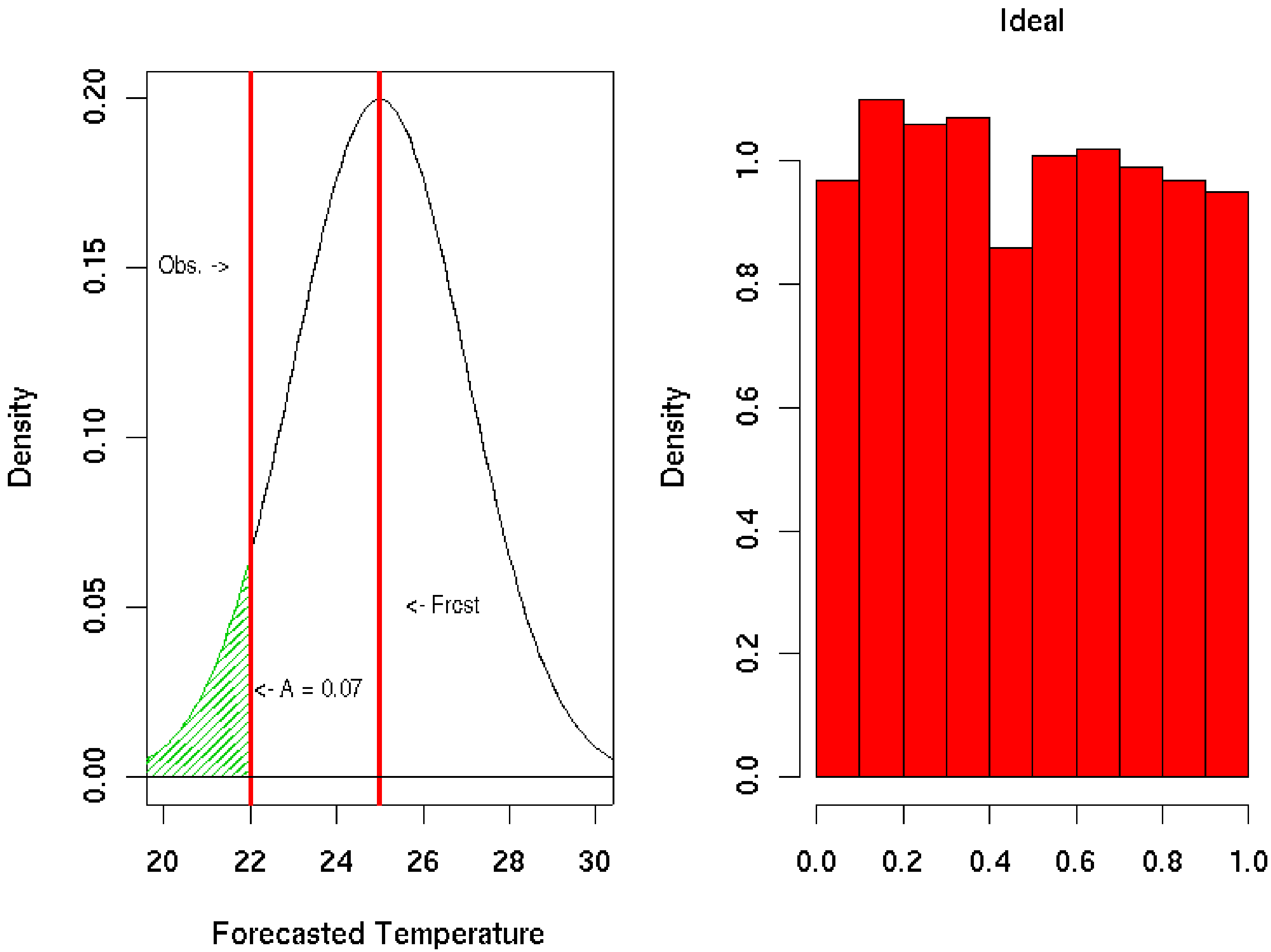


Rank histograms are a way to examine the calibration of an ensemble

Verifying a continuous expression of a distribution (i.e. normal, Poisson, beta)

- Probability of any observation occurring is on [0,1] interval.
- Probability Integral Transformed (PIT) - fancy word for how likely is a given forecast
- Still create a rank histogram using bins of probability of observed events.

Verifying a distribution forecast



Warnings about rank histograms

- Assume all samples come from the same climatology!
- A flat rank histogram can be derived by combining forecasts with offsetting biases
- See Hamill, T. M., and J. Juras, 2006: Measuring forecast skill: is it real skill or is it the varying climatology? *Quart. J. Royal Meteor. Soc.*, Jan 2007 issue
- Techniques exist for evaluating "flatness", but they mostly require much data.

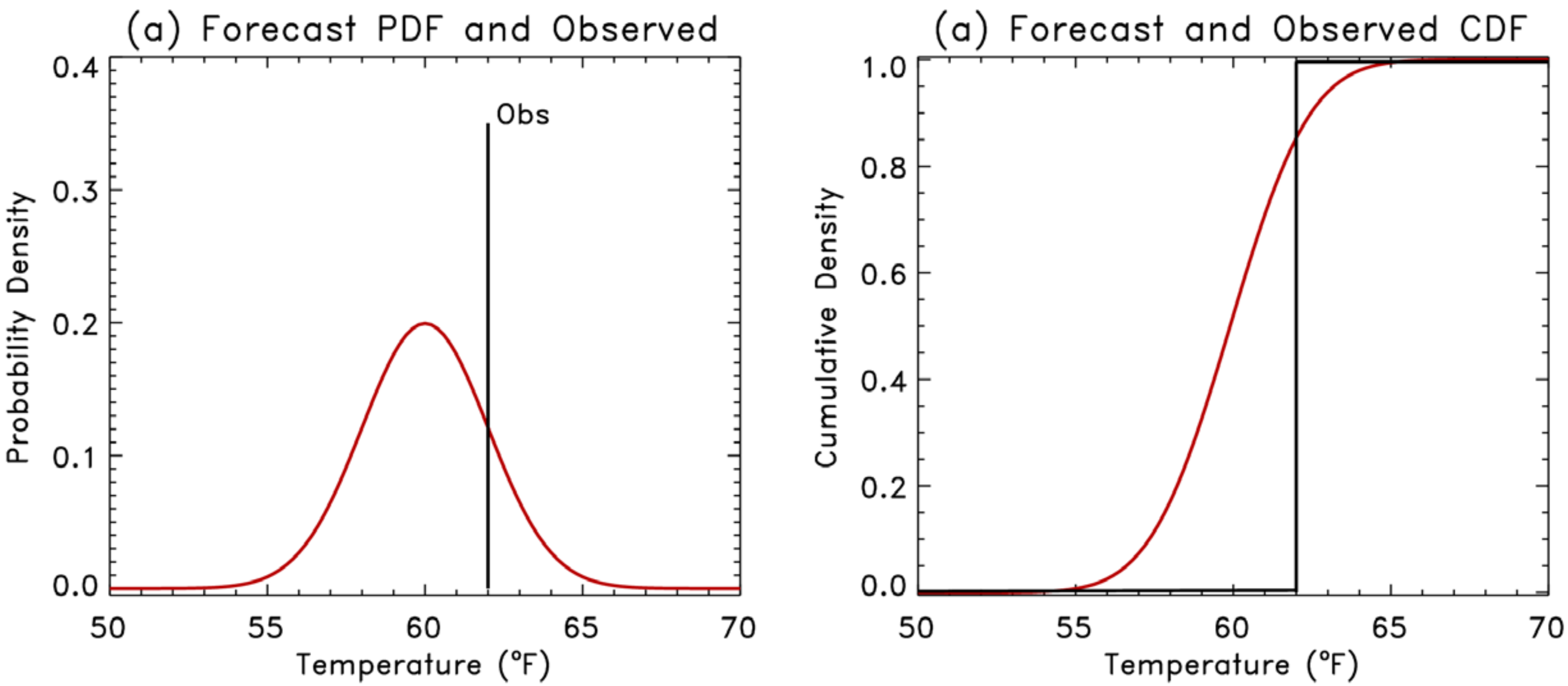
Continuous and discrete rank probability scores

- Measures of accuracy for
  - Multiple category forecasts (e.g., precipitation type)

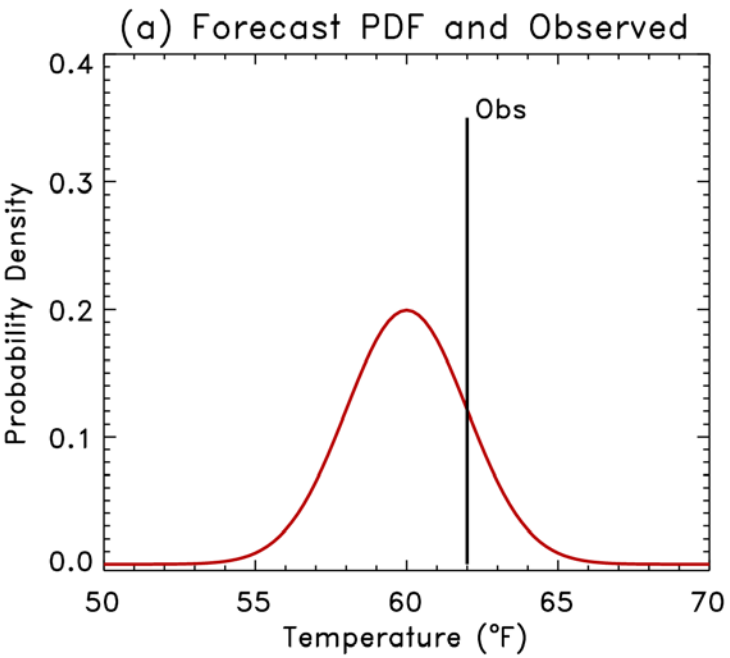
  Rank Probability Score (RPS)
  - Continuous distributions (e.g., ensemble distribution)

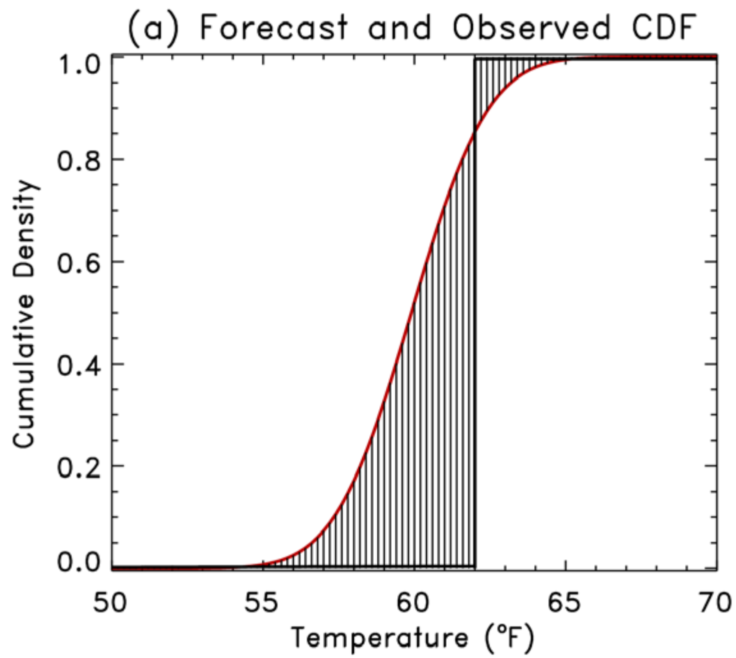
  Continuous Ranked Probability Score (CRPS)
- Relates to Brier score – for a forecast of a binary event, the RPS score is equivalent to the Brier score.

Rank Probability Scores



A good RPS score minimizes area





Ignorance score (for multi-category or ensemble forecasts)

- A "local" score

$$\mathrm{IS} = \frac{1}{n}\sum_{i=1}^{n} \log_2(\widehat{p}_{t,k^*(t)})$$

- $k^*(t)$ is the category that actually was observed at time t
- Based on information theory
- Only rewards forecasts with some probability in "correct" category
- Perfect score: 0 [i.e., $\log_2(1) = 0$]

Final comments

- Know how and why your ensemble is being created.
- Use a combination of graphics and scores.
- Areas of more research
  - Verification of spatial forecasts
  - Additional intuitive measures of performance for probability and ensemble forecasts.

Measure	**Attribute evaluated**	**Comments**
Probability forecasts		
Brier score	Accuracy	Based on squared error
Resolution	Resolution (resolving different categories)	Compares forecast category climatologies to overall climatology
Reliability	Calibration	
Skill score	Skill	Skill involves *comparison* of forecasts
Sharpness measure	Sharpness	Only considers distribution of forecasts
Relative Operating Characteristic (ROC)	Discrimination	Ignores calibration
C/L Value	Value	Ignores calibration
Ensemble distribution		
Rank histogram	Calibration	Can be misleading
Spread-skill	Calibration	Difficult to achieve
CRPS	Accuracy	Squared difference between forecast and observed distributions Analogous to MAE in limit
log p score	Accuracy	Local score, rewards for correct category; infinite if observed category has 0 density

Useful references

- **Good overall references** for forecast verification:
  - (1) Wilks, D.S., 2006: *Statistical Methods in the Atmospheric Sciences (2nd Ed).* Academic Press, 627 pp.
  - (2) WMO Verification working group forecast verification web page, http://www.cawcr.gov.au/projects/verification/
  - (3) Jolliffe and Stephenson Book: Jolliffe, I.T., and D.B. Stephenson, 2012: *Forecast Verification. A Practitioner's Guide in Atmospheric Science*., 2nd Edition, Wiley and Sons Ltd.
- **Verification tutorial – Eumetcal** (http://www.eumetcal.org/-learning-modules-)
- **Rank histograms**: Hamill, T. M., 2001: Interpretation of rank histograms for verifying ensemble forecasts. *Mon. Wea. Rev*., **129**, 550-560.
- **Spread-skill relationships**: Whitaker, J.S., and A. F. Loughe, 1998: The relationship between ensemble spread and ensemble mean skill. *Mon. Wea. Rev*., **126**, 3292-3302.
- **Brier score, continuous ranked probability score, reliability diagrams**: Wilks text again.
- **Relative operating characteristic**: Harvey, L. O., Jr, and others, 1992: The application of signal detection theory to weather forecasting behavior. *Mon. Wea. Rev*., **120**, 863-883.
- **Economic value diagrams**:
  - (1)Richardson, D. S., 2000: Skill and relative economic value of the ECMWF ensemble prediction system. *Quart. J. Royal Meteor. Soc*., **126**, 649-667.
  - (2) Zhu, Y, and others, 2002: The economic value of ensemble-based weather forecasts. *Bull. Amer. Meteor. Soc*., **83**, 73-83.
- **Overestimating skill**: Hamill, T. M., and J. Juras, 2006: Measuring forecast skill: is it real skill or is it the varying climatology? *Quart. J. Royal Meteor. Soc.*, Jan 2007 issue. http://tinyurl.com/kxtct

Reliability Diagram Exercise

Probabilities underforecast

No resolution

Essentially no skill

Over-resolved forecast

Perfect forecast

Typical categorical forecast