Intro to Numerical Model Evaluation and Ensembles at the Joint Numerical Testbed

Presented by
Tressa L. Fowler

DTC
Developmental Testbed Center

- The main goals of the JNT are to test and evaluate numerical weather prediction (NWP) systems to provide meaningful information about forecast performance to operational decision makers and to provide the research community with support in their development of these systems.

Community software

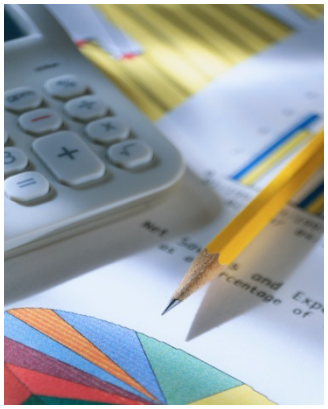
- General model verification issues
  - Matching observations to models
  - Evaluate via statistics
- Ensemble verification
- MET capabilities and examples

What is verification?

- Verification is the process of comparing forecasts to relevant observations
  - Verification is one aspect of measuring forecast ***goodness***
- Verification measures the ***quality*** of forecasts (as opposed to their ***value***)
- For many purposes a more appropriate term is "***evaluation***"

Why verify?

- Purposes of verification (traditional definition)



  - Administrative purpose
    - Monitoring performance
    - Choice of model or model configuration (has the model improved?)

  - Scientific purpose
    - Identifying and correcting model flaws
    - Forecast improvement

  - Economic purpose
    - Improved decision making
    - "Feeding" decision models or decision support systems

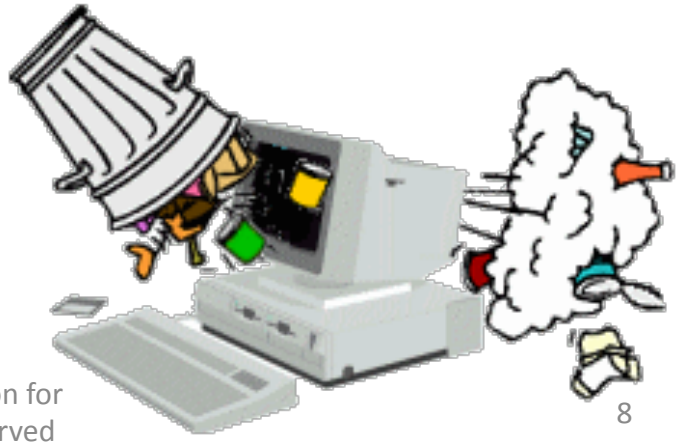
Identifying verification goals

What *questions* do we want to answer?

- Examples:
  - ✓In what locations does the model have the best performance?
  - ✓Are there regimes in which the forecasts are better or worse?
  - ✓Is the probability forecast well calibrated (i.e., reliable)?
  - ✓Do the forecasts correctly capture the natural variability of the weather?

Useful verification

- Using MET is the easy part, scientifically speaking.
- Good verification depends mostly on what you do before and after MET.
  - What do you want to know?
  - Good forecasts.
  - Good observations.
  - Well matched.
  - Appropriate selection of methods
  - Thorough and correct interpretation of results.



Observations

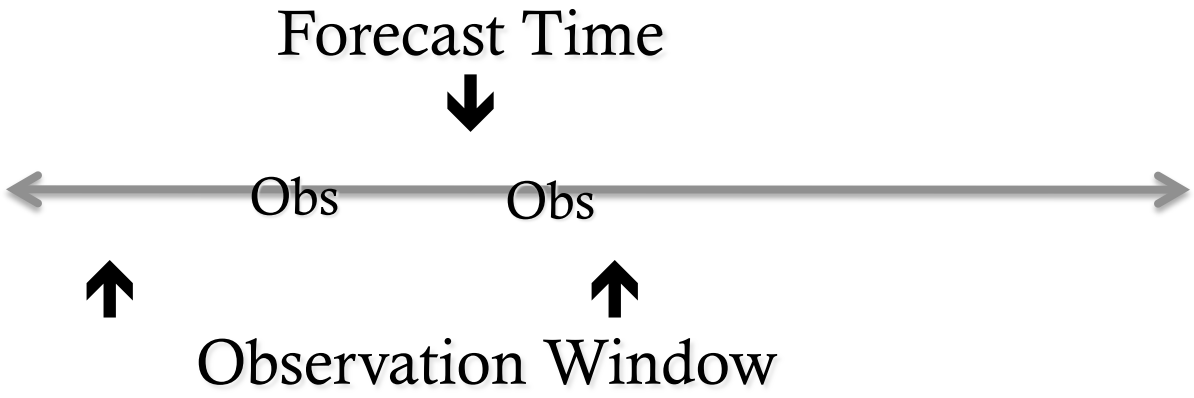
Observations should represent the ***event*** being forecast, including the

- Element (e.g., temperature, precipitation)
- Temporal resolution
- Spatial resolution and representation
- Thresholds, categories, etc.

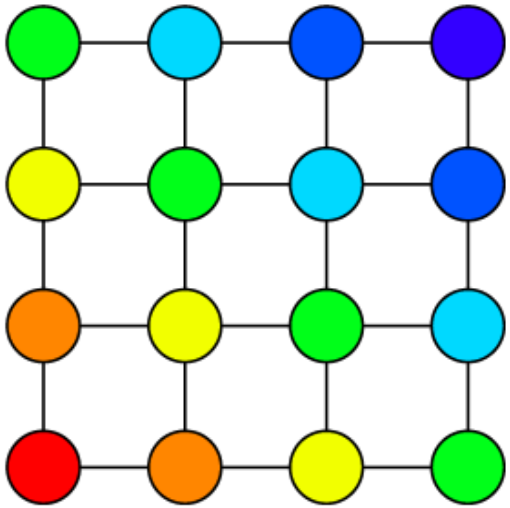
- Good observations are
  - Independent of the forecasts
  - Largely independent of each other
  - Inaccuracy, variability, and biases are known and accounted for.

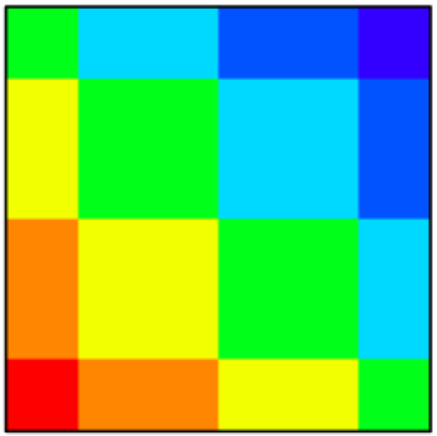
Time

- Forecasts and observations are probably not at the same time.
- They are often 'matched' by using a time window.

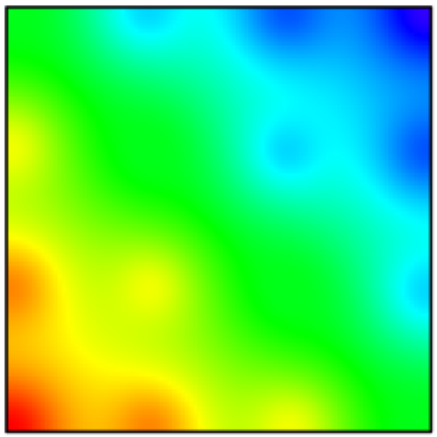


Matching observations to forecasts in space – interpolation examples

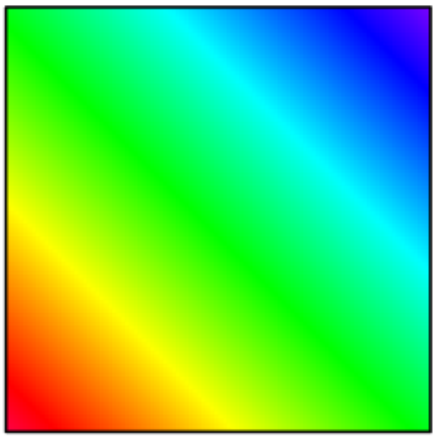




Nearest Neighbor



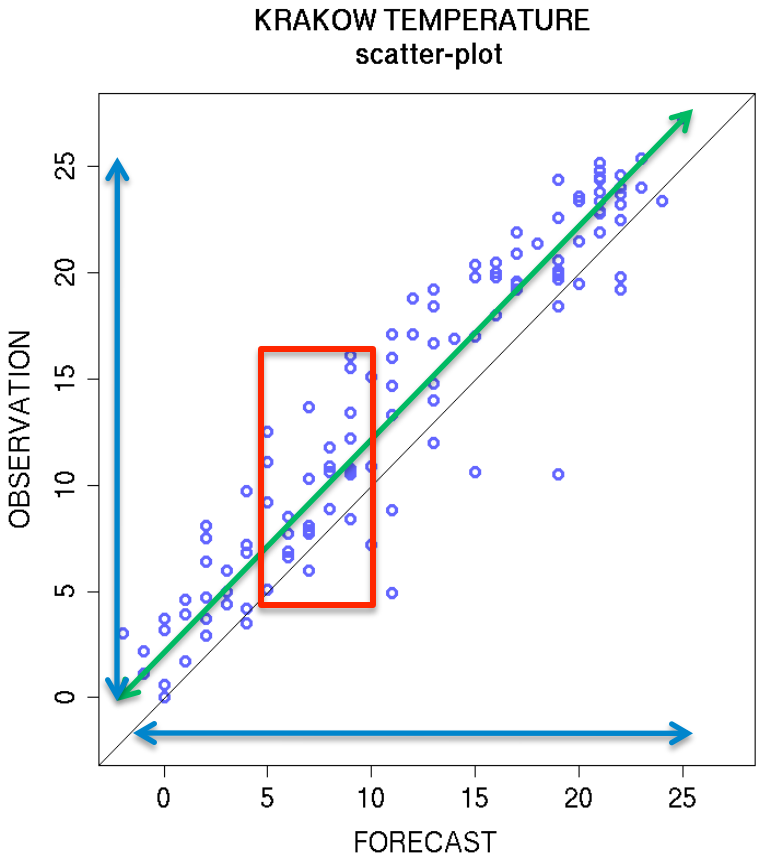
Distance Weighted Mean



Least Squares

Verification attribute examples

- Bias
  - (Marginal distributions)
- Correlation
  - Overall association (Joint distribution)
- Accuracy
  - Differences (Joint distribution)
- Calibration
  - Measures conditional bias (Conditional distributions)
- Discrimination
  - Degree to which forecasts discriminate between different observations (Conditional distribution)



Joint : The probability of two events in conjunction.

Pr (Tornado forecast AND Tornado observed) = 30 / 2800 = 0.01

Conditional : The probability of one variable given that the second is already determined.

Pr (Tornado Observed | Tornado Fcst) = 30/50 = 0.60

Marginal : The probability of one variable without regard to the other.

Tornado forecast	Tornado Observed		
	yes	no	Total fc
yes	30	70	100
no	20	2680	2700
Total obs	50	2750	2800

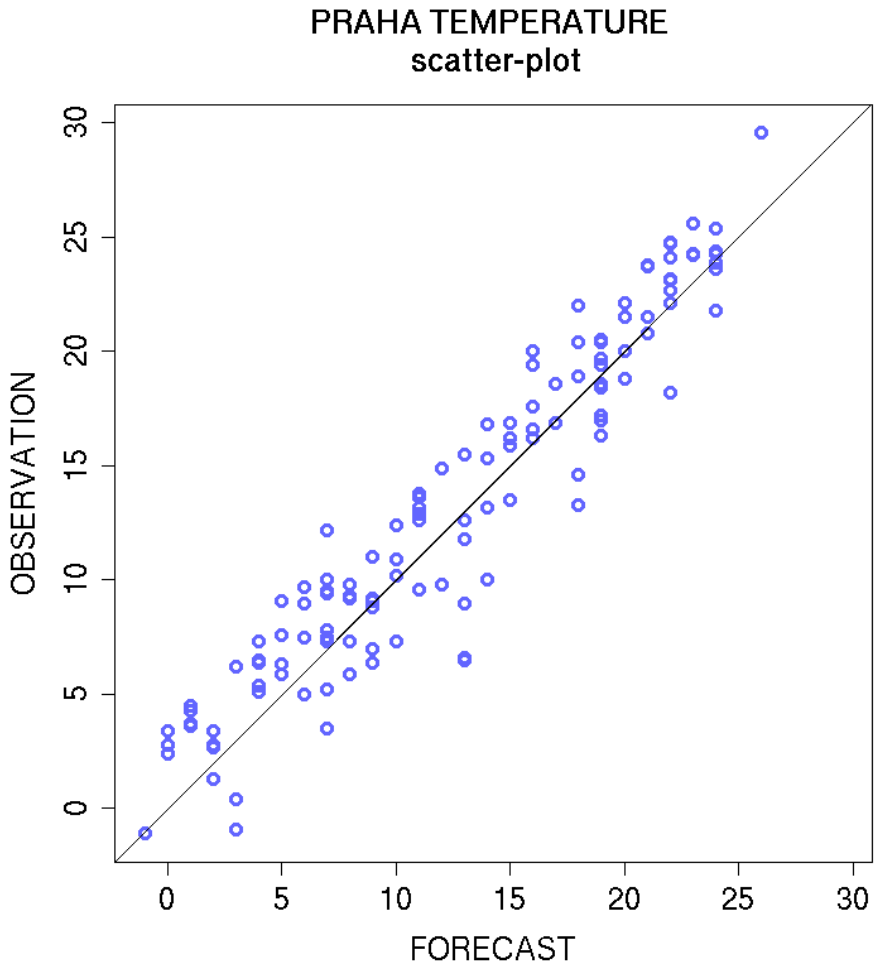
Pr(Yes Forecast) = 100/2800 = 0.04
Pr(Yes Obs) = 50 / 2800 = 0.02

Exploratory methods: joint distribution

Scatter-plot: plot of observation versus forecast values

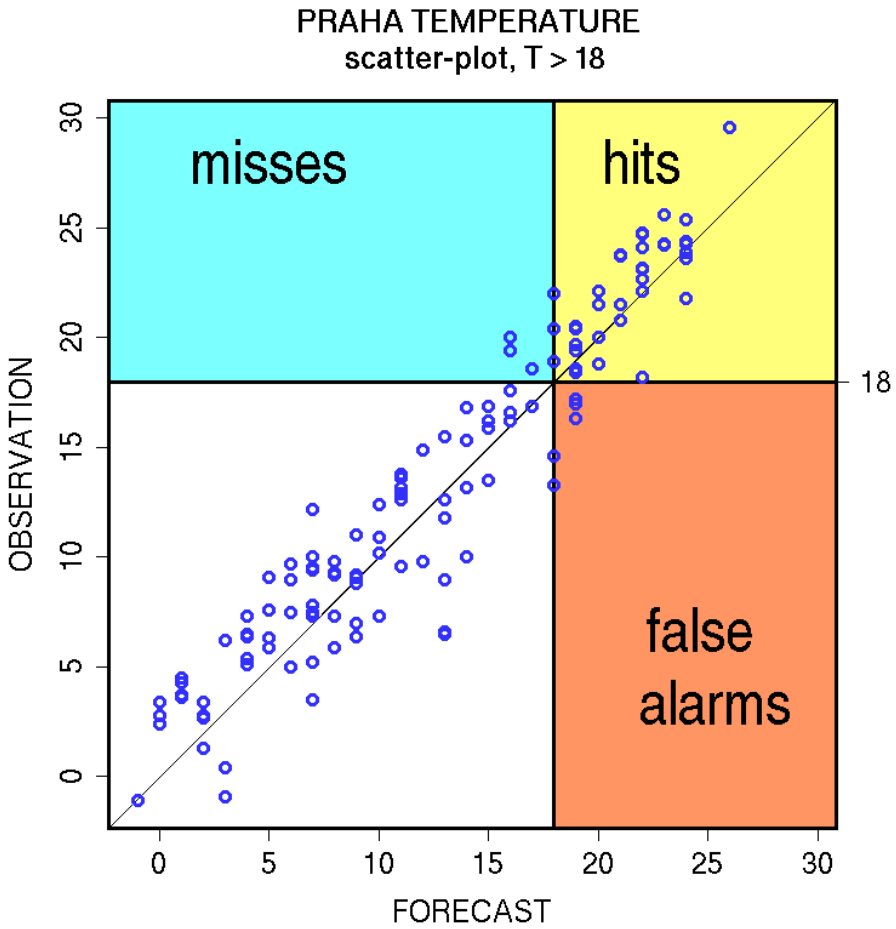
Perfect forecast = obs, points should be on the 45^{o} diagonal

Provides information on: bias, outliers, error magnitude, linear association, peculiar behaviours in extremes, misses and false alarms (link to contingency table)

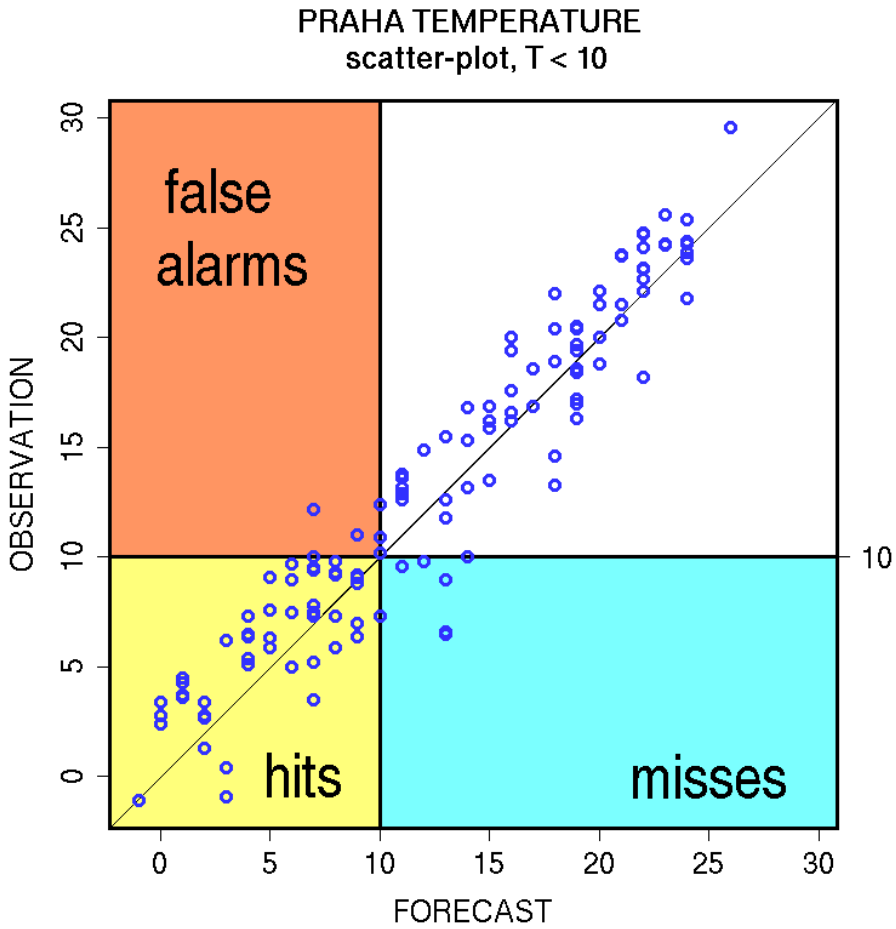


Scatter-plot and Contingency Table

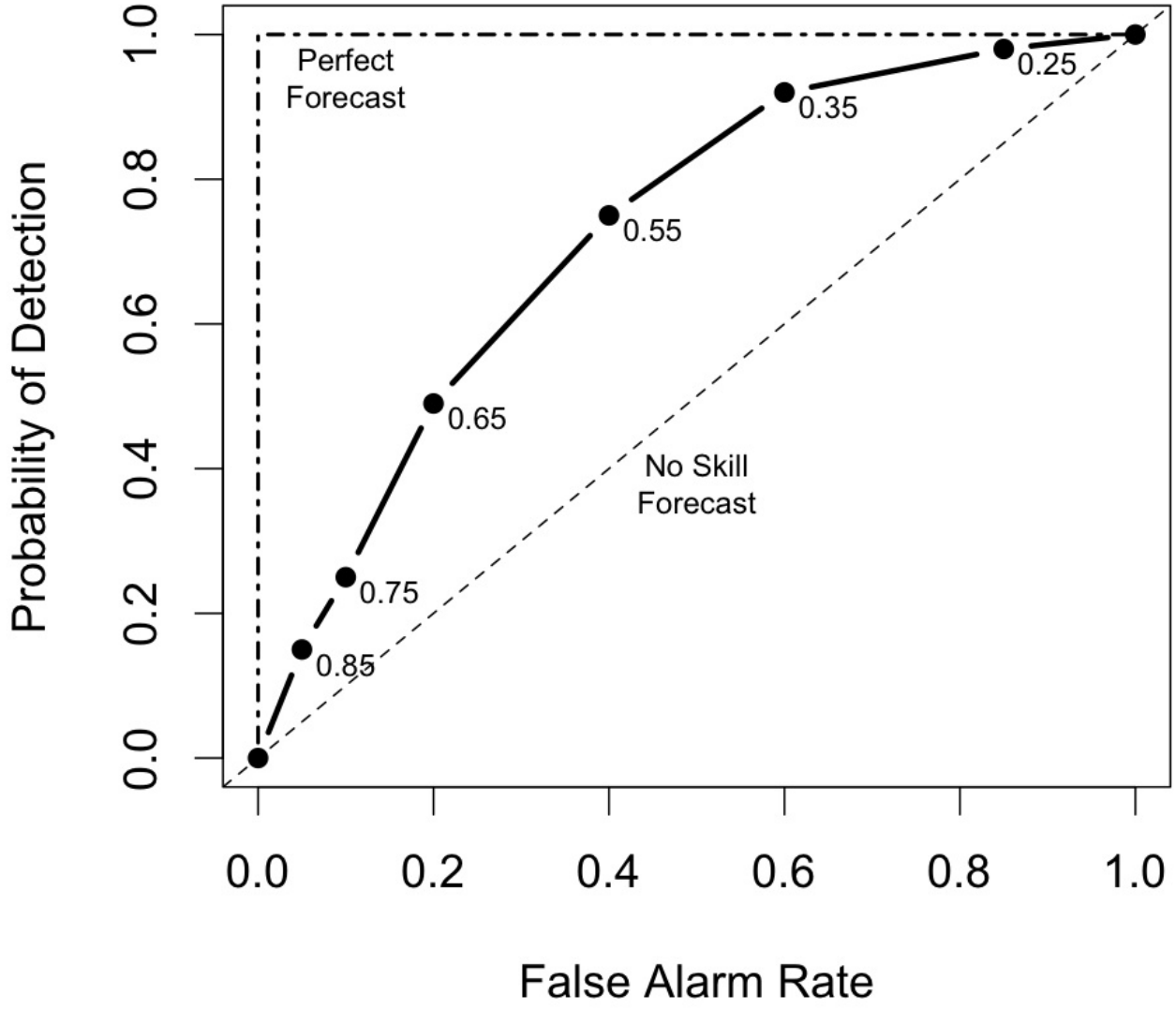
Does the forecast detect correctly temperatures above 18 degrees ?



Does the forecast detect correctly temperatures below 10 degrees ?



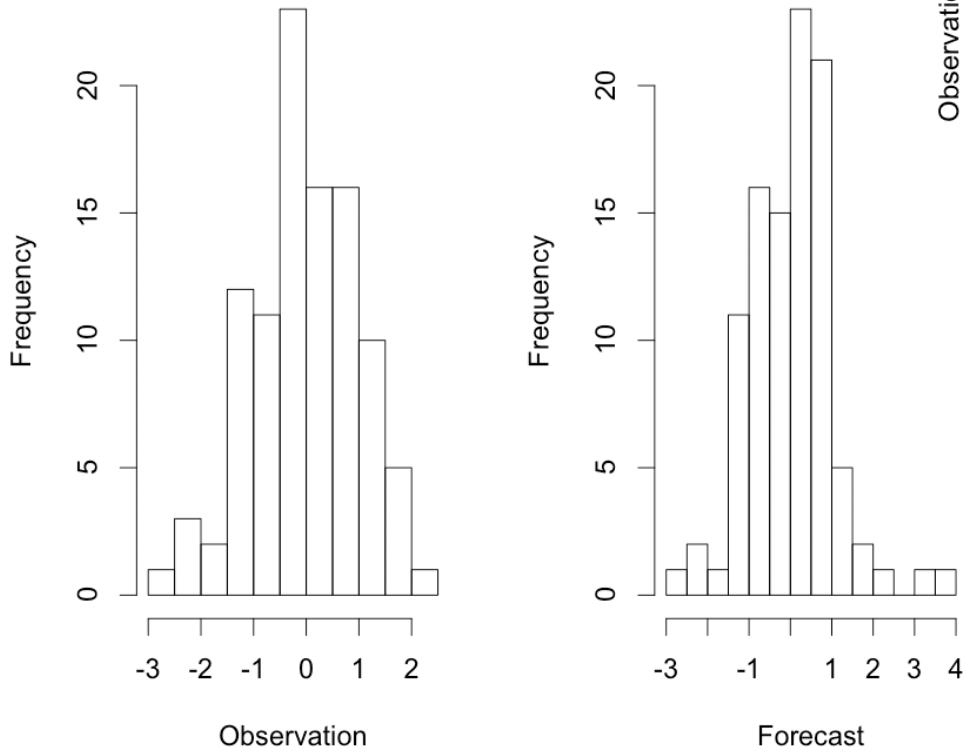
Example Receiver Operating Characteristic Plot

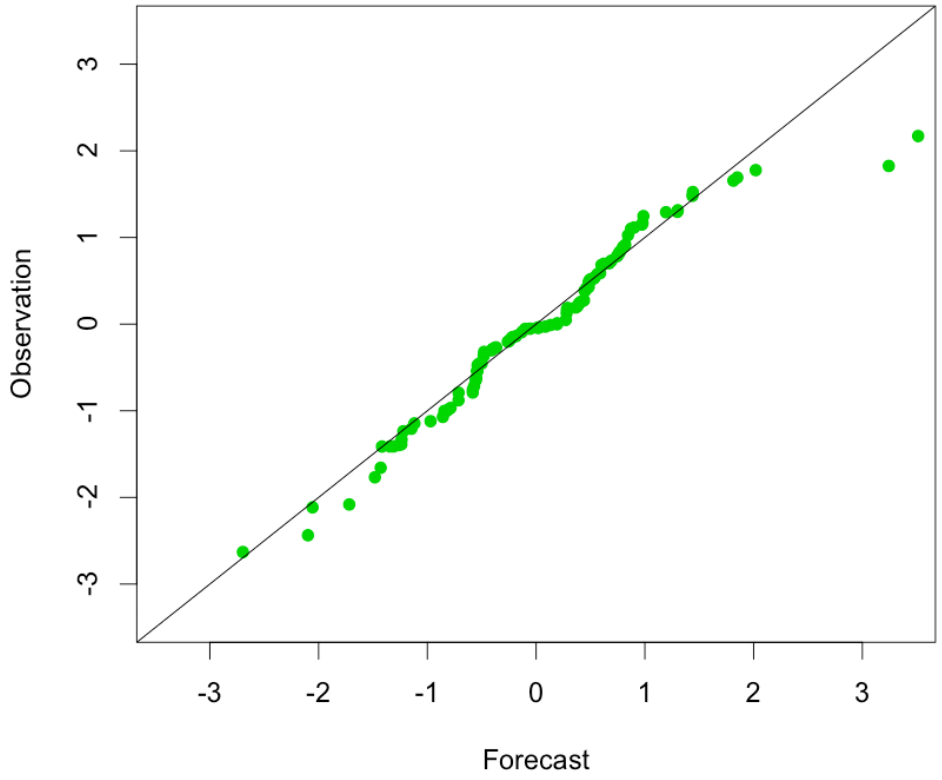


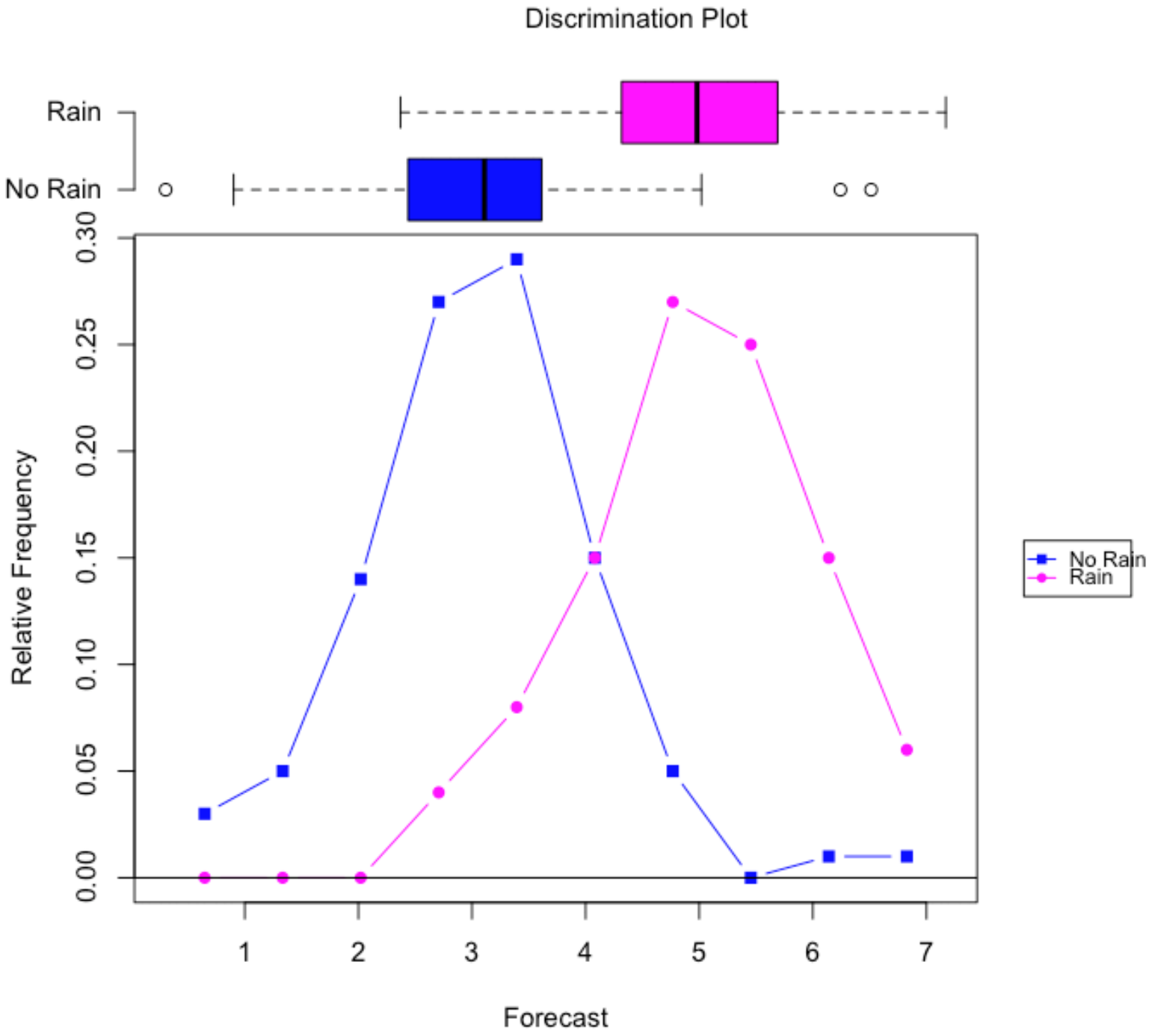
Exploratory methods: marginal distribution

Quantile-quantile plots:

OBS quantile versus the corresponding FRCS quantile






Discrimination Plot
Rain
No Rain
Relative Frequency
0.00
0.05
0.10
0.15
0.20
0.25
0.30
1
2
3
4
5
6
7
Forecast
No Rain
Rain

Exploratory methods: marginal distributions

Visual comparison:
Histograms, box-plots, …

Summary statistics:

- ***Location:***

$$\text{mean} = \overline{X} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

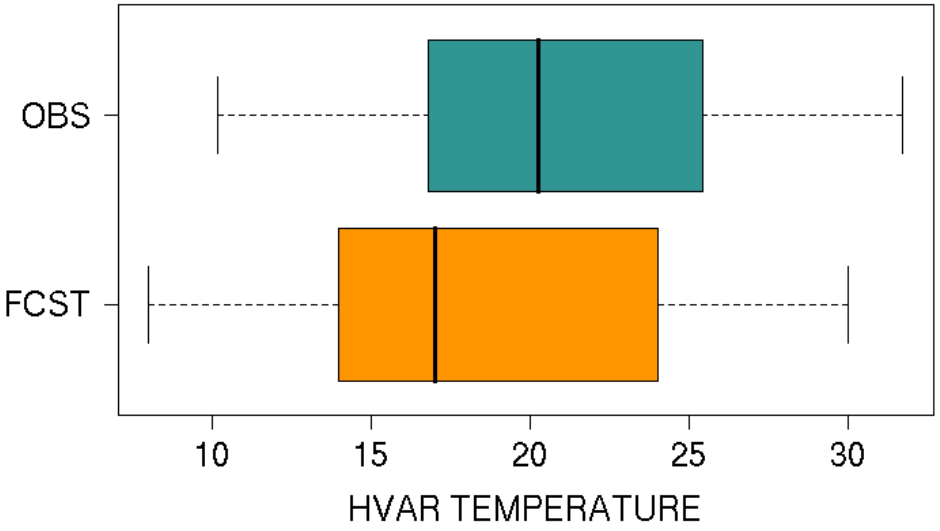
$$\text{median} = q_{0.5}$$

- ***Spread:***

$$\text{st dev} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \overline{X}\right)^2}$$

$$\text{Inter Quartile Range} =$$

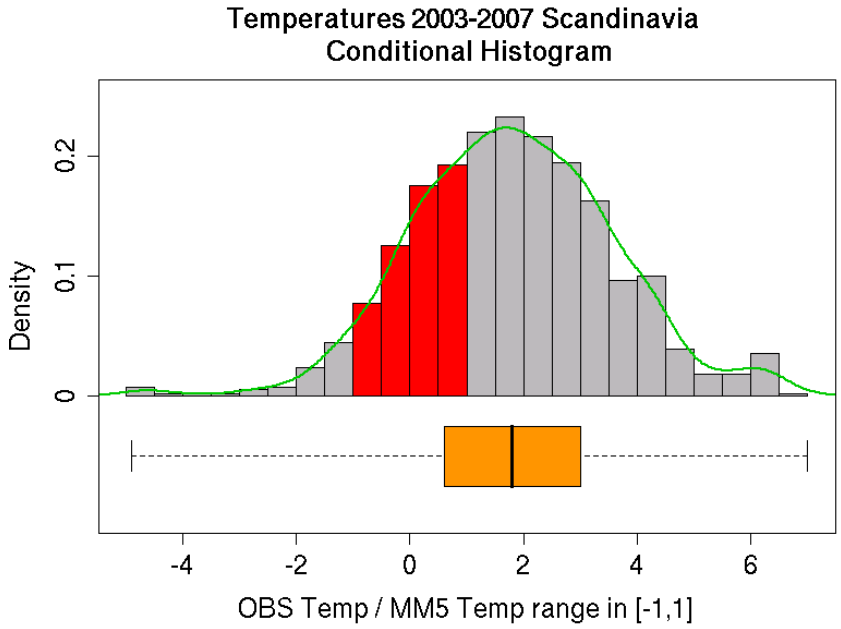
$$\text{IQR} = q_{0.75} - q_{0.25}$$

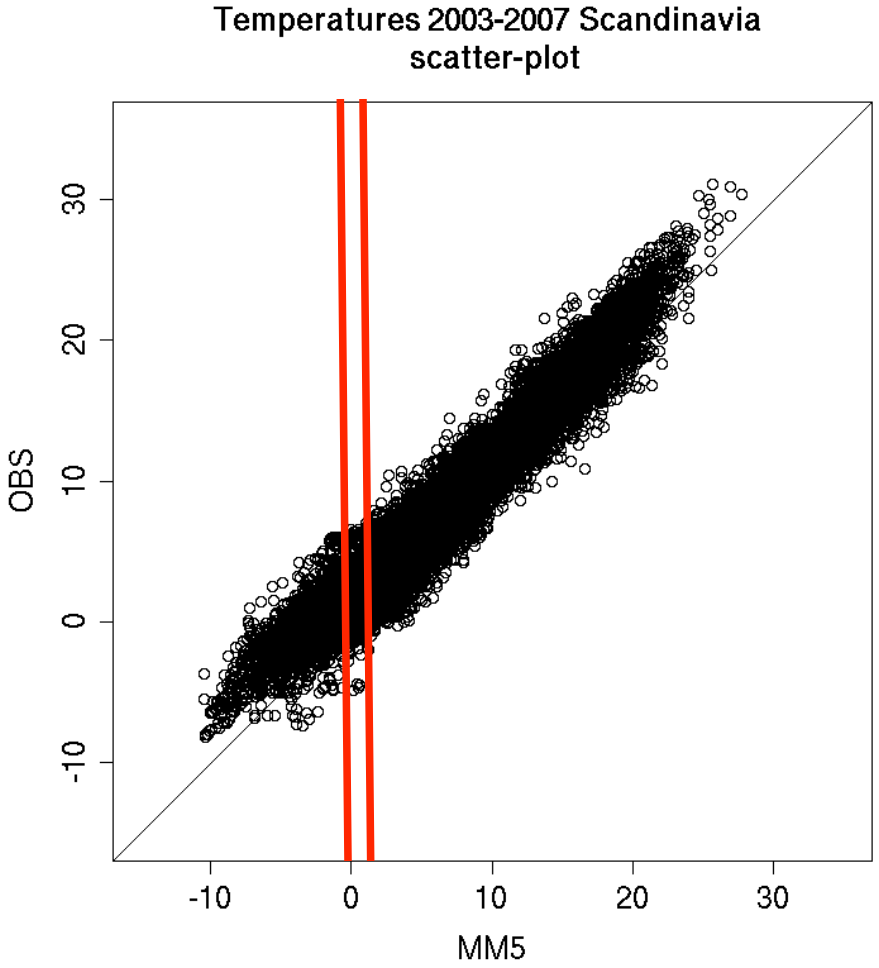


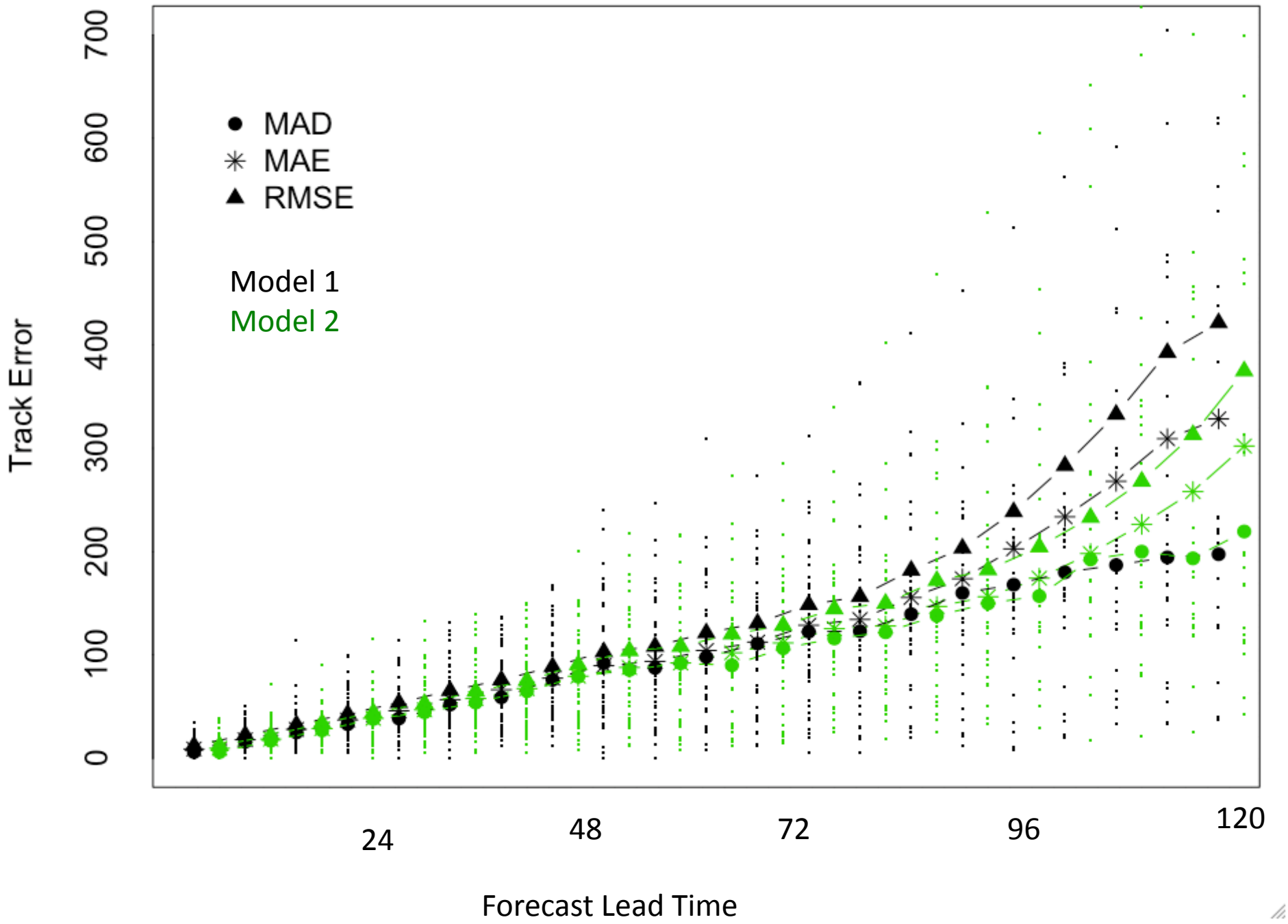
	MEAN	MEDIAN	STDEV	IQR
OBS	20.71	20.25	5.18	8.52
FRCS	18.62	17.00	5.99	9.75

Exploratory methods: conditional distributions

Conditional histogram and conditional box-plot






MAD
MAE
RMSE
Model 1
Model 2
Track Error
0
100
200
300
400
500
600
700
24
48
72
96
120
Forecast Lead Time

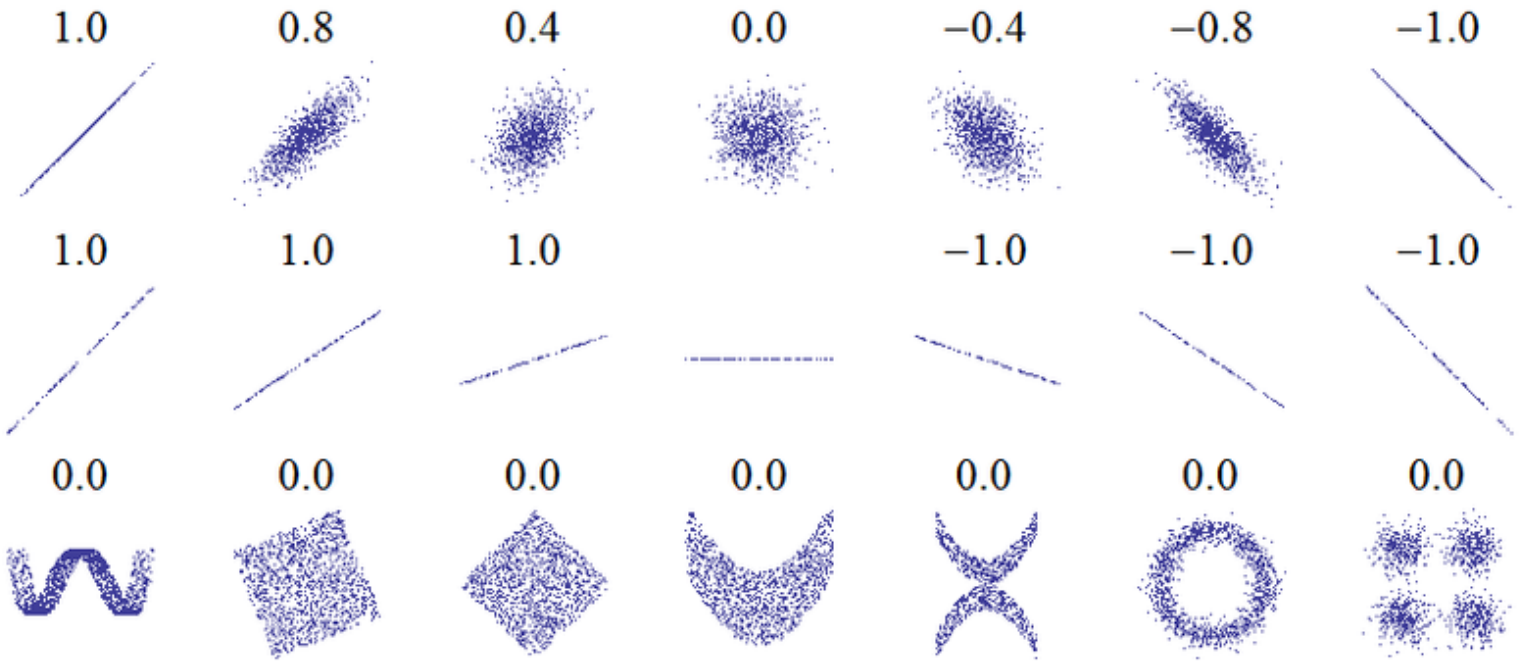
Scores for continuous forecasts

Simplest overall measure of performance:

Correlation coefficient

$$\rho_{fx} = \frac{Cov(f,x)}{\sqrt{Var(f)Var(x)}}$$

$$r_{fx} = \frac{\sum_{i=1}^{n}(f_i - \bar{f})(x_i - \bar{x})}{(n-1)s_f s_x}$$



MSE and bias correction

$$MSE = \left(\overline{f} - \overline{o}\right)^2 + s_f^2 + s_o^2 - 2s_f s_o r_{fo}$$

$$MSE = ME^2 + \mathrm{var}(f - o)$$

- MSE is the sum of the squared bias and the variance. So ↑ bias = ↑ MSE

Continuous Scores of **Ranks**

Problem: Continuous scores sensitive to large values or non robust.

Solution: Use the **ranks** of the variable, rather than its actual values.

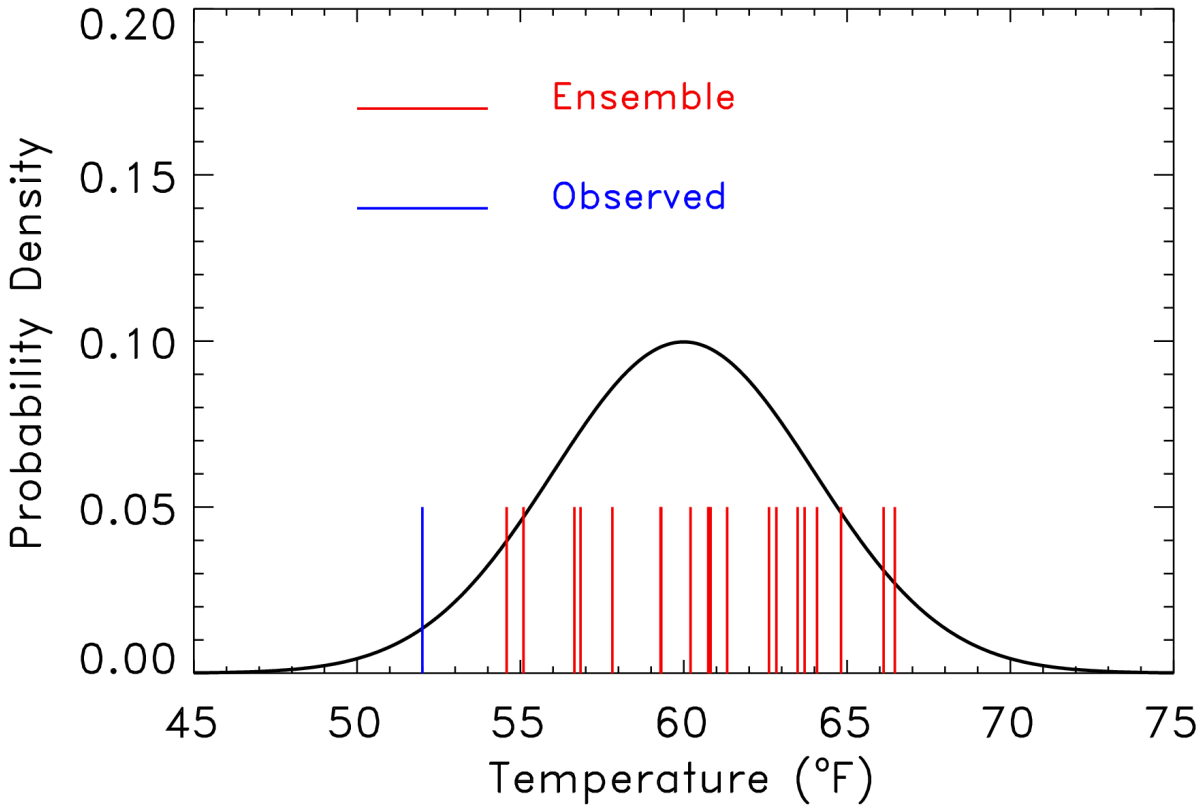
Temp °C	27.4	21.7	24.2	23.1	19.8	25.5	24.6	22.3
rank	8	2	5	4	1	7	6	3

The value-to-rank transformation:

- diminish effects due to large values
- transform distribution to a Uniform distribution
- remove bias

Rank correlation is the most common.

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

Verifying a probabilistic forecast

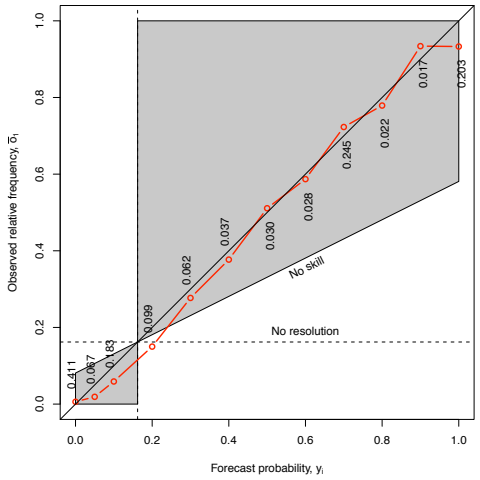
- You cannot verify a probabilistic forecast with a single observation.
- The more data you have for verification, (as with other statistics) the more certain you are.
- Rare events (low probability) require more data to verify.
- These comments refer to probabilistic forecasts developed by methods other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

Evaluate each member as a separate, deterministic forecast

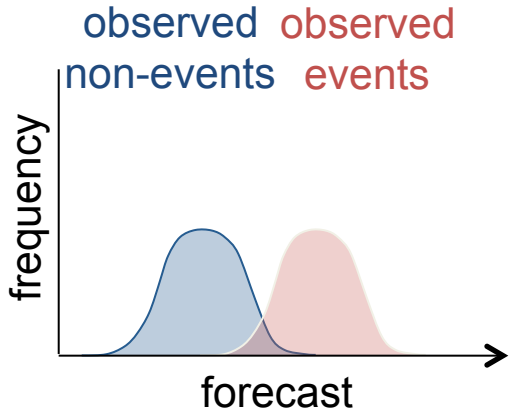
- Why? Because it is easy and important
  - If members are unique, it might provide useful diagnostics.
  - If members are bias, verification statistics might be skewed.
  - If members have different levels of bias, should you calibrate?
  - Do these results conform to your understanding of how the ensemble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

Properties of a perfect probabilistic forecast of a binary event.

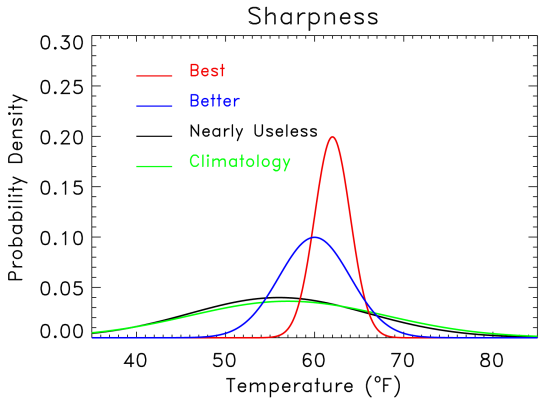
Reliability



Resolution



Sharpness



Introducing the attribute diagram!

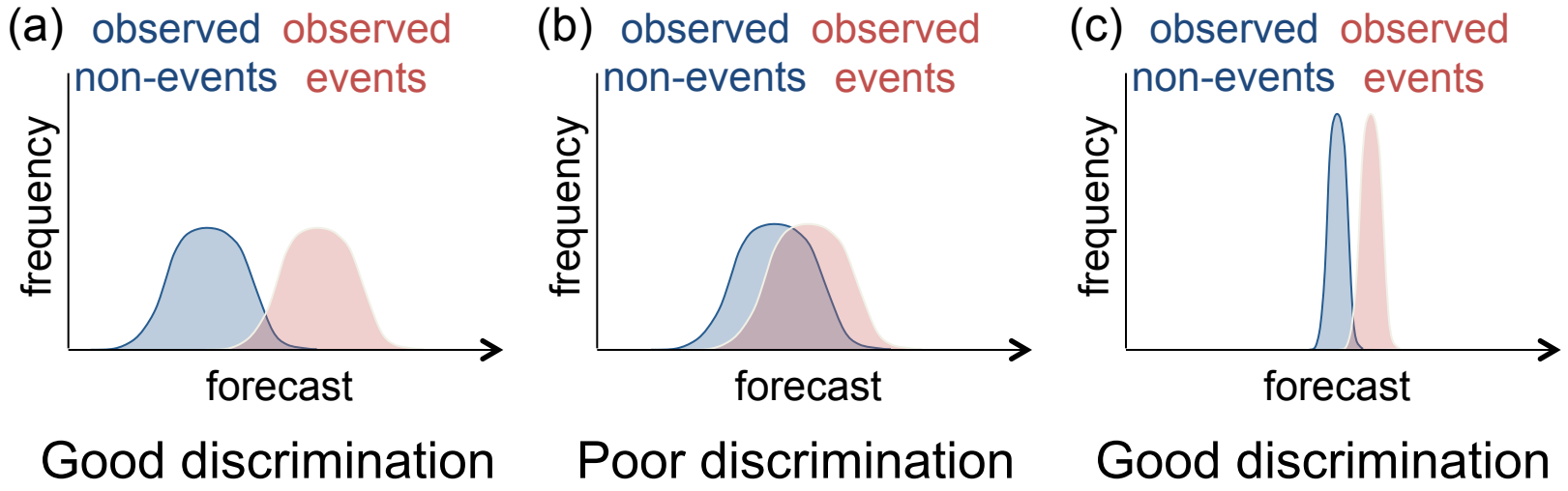
(close relative to the reliability diagram)

- 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

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



Brier score

- Appropriate for probabilities of a binary event occurring.
- Analogous to mean square error.
- Can be decomposed to resolution, reliability and uncertainty components.
- Geometrically relates to attribute diagram.
- Fair comparisons require common sample climatology.

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

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.

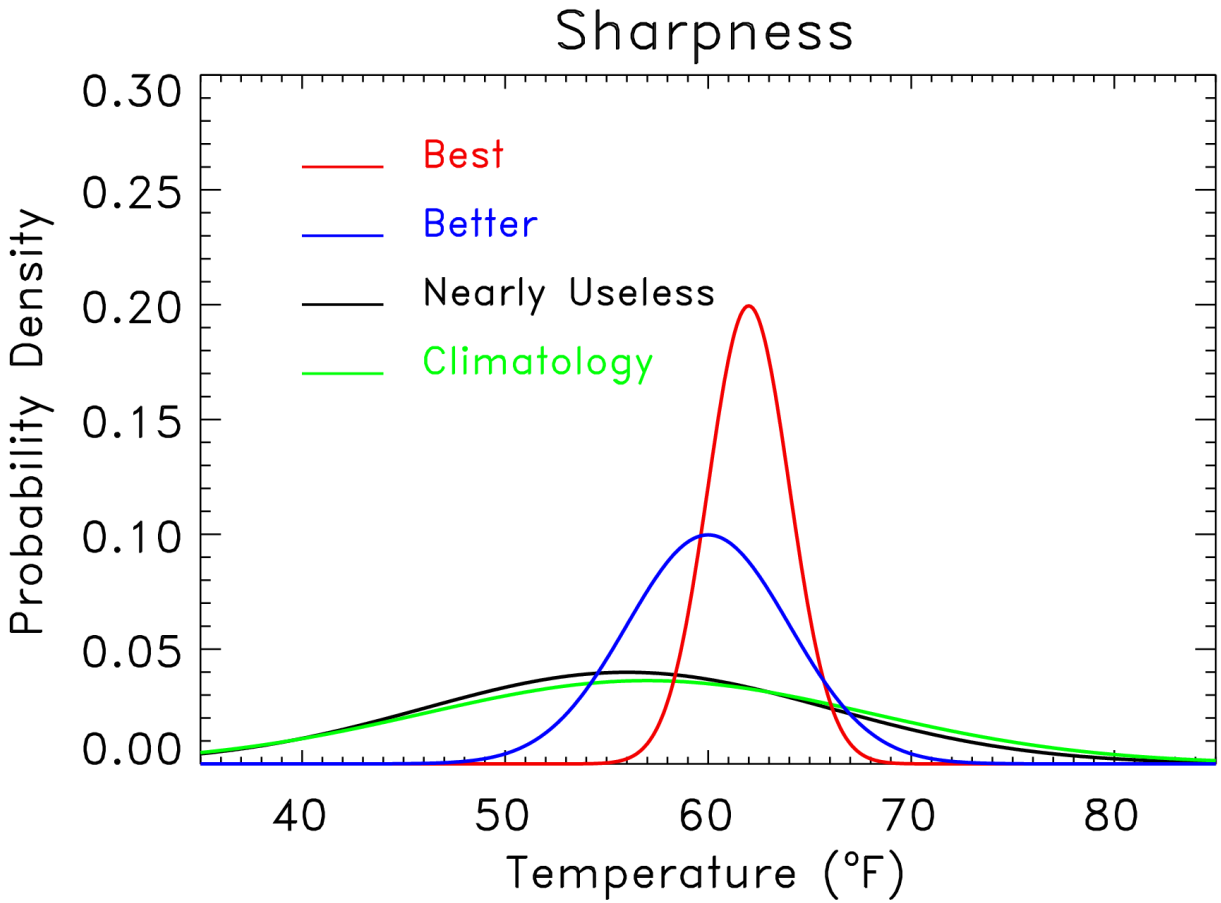
uncertainty

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.

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.

$$\mathrm{BS} = \frac{1}{n}\sum_{i=1}^{I} N_i\left(y_i - \overline{o}_i\right)^2 - \frac{1}{n}\sum_{i=1}^{I} N_i\left(\overline{o}_i - \overline{o}\right)^2 + \overline{o}\left(1 - \overline{o}\right)$$

("reliability") ("resolution") ("uncertainty")

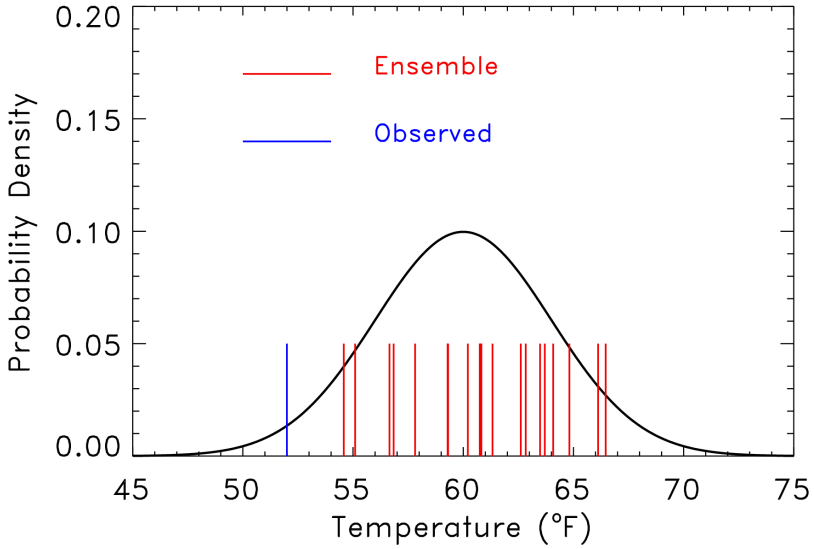
Forecasts of a full distribution

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

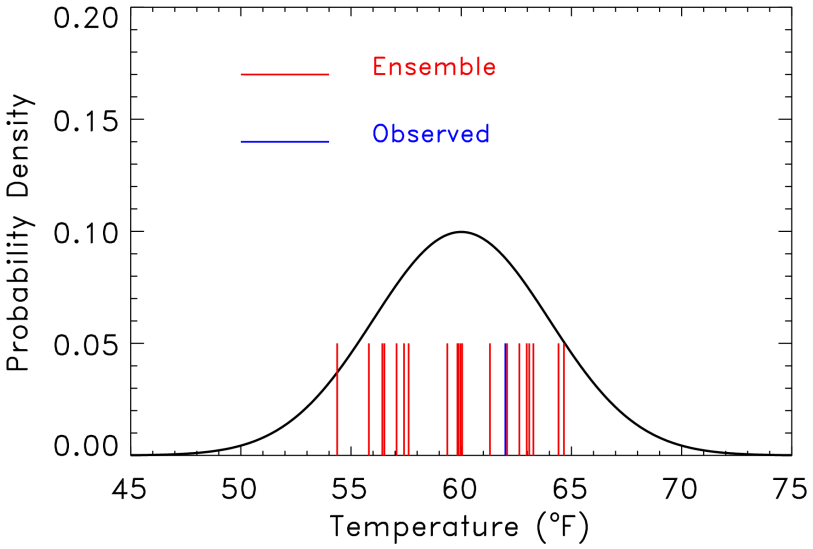
Assuming the forecast is reliable (calibrated)

- 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.
- Huh?

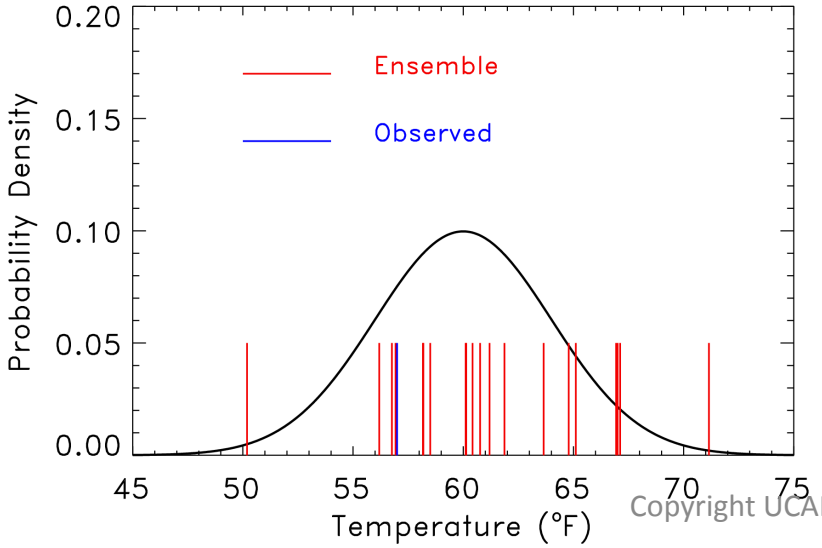
Rank 1 of 21



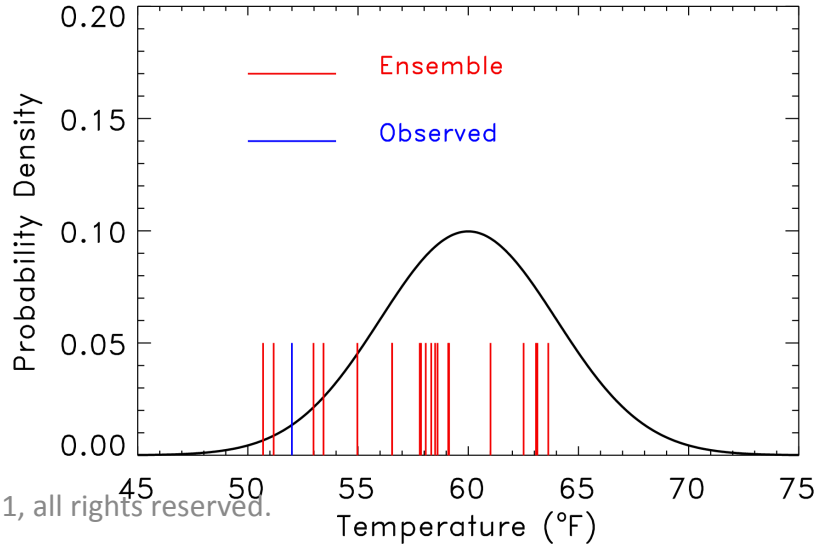
Rank 14 of 21



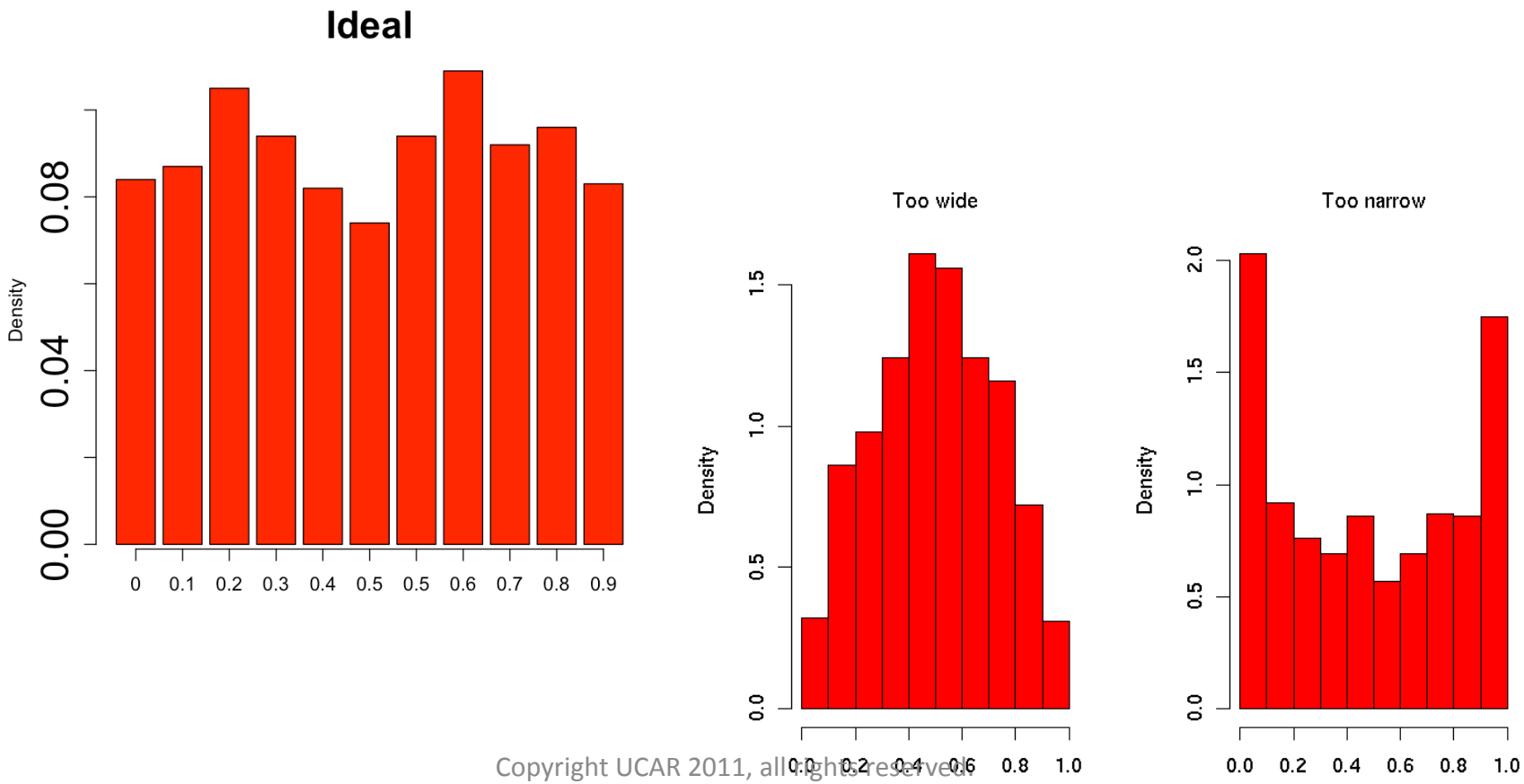
Rank 5 of 21



Rank 3 of 21



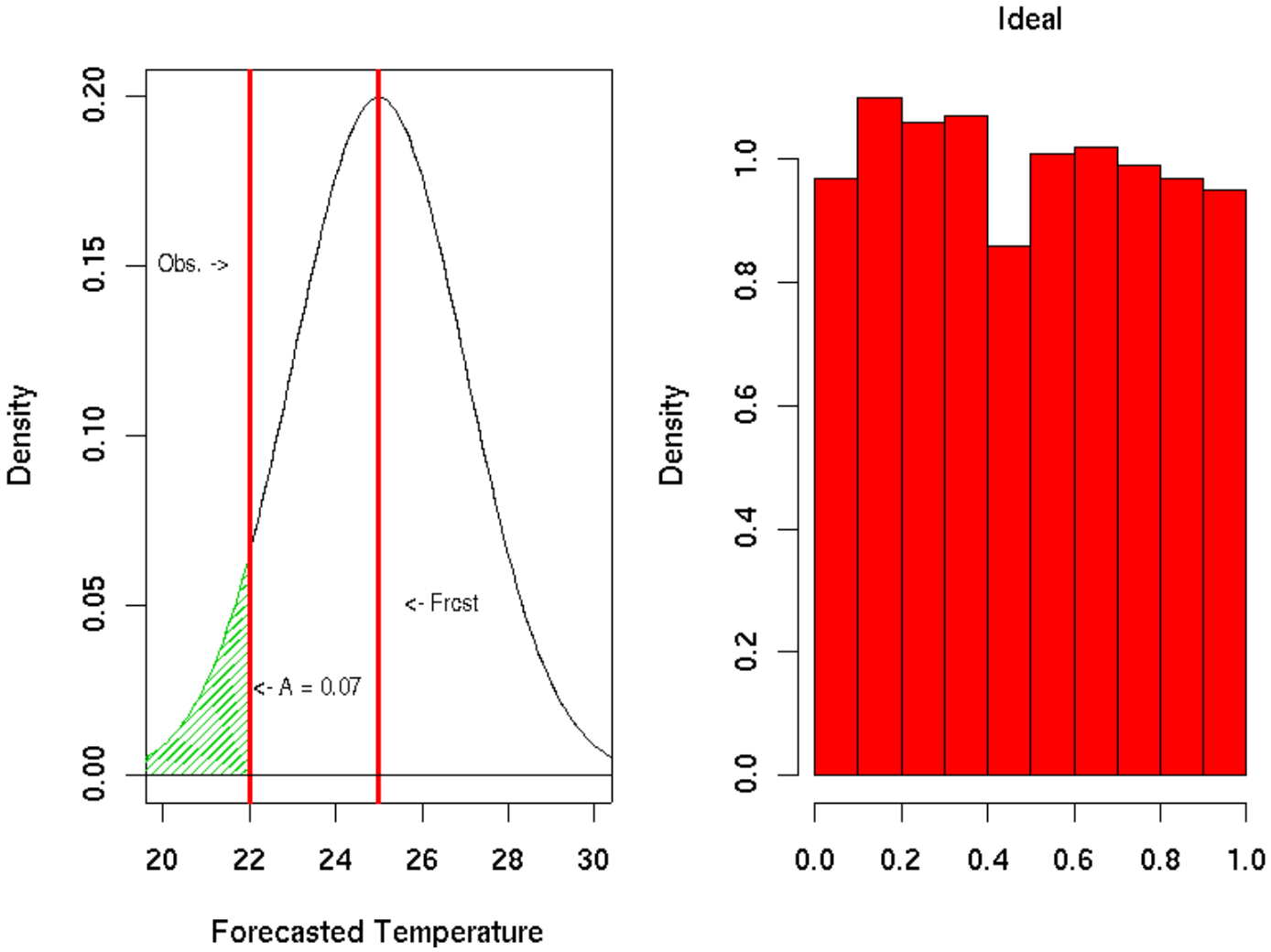
Rank Histograms



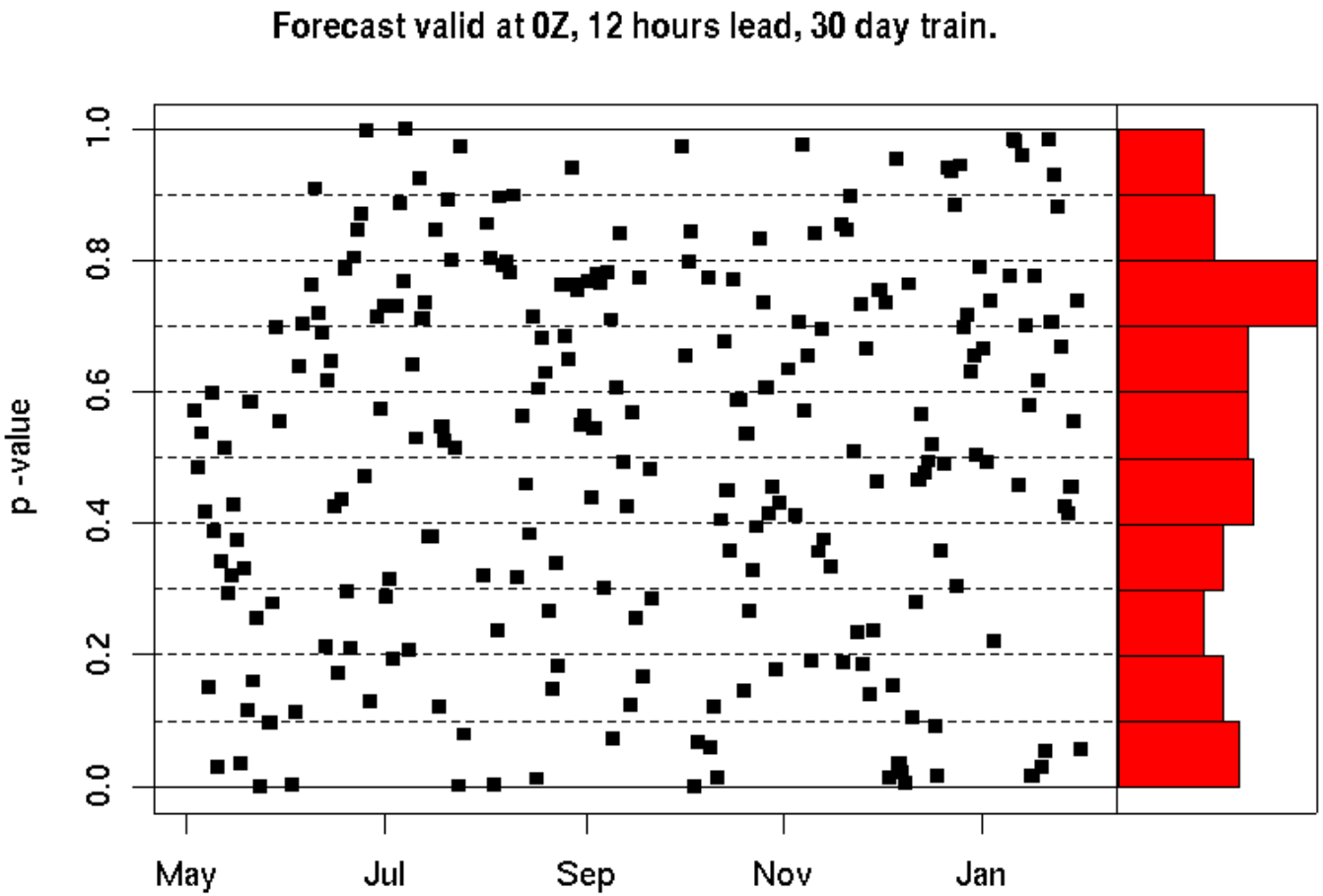
Verifying a continuous expression of a distribution (i.e. normal, Poisson, beta)

- 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



Evaluate order of probabilistic forecasts



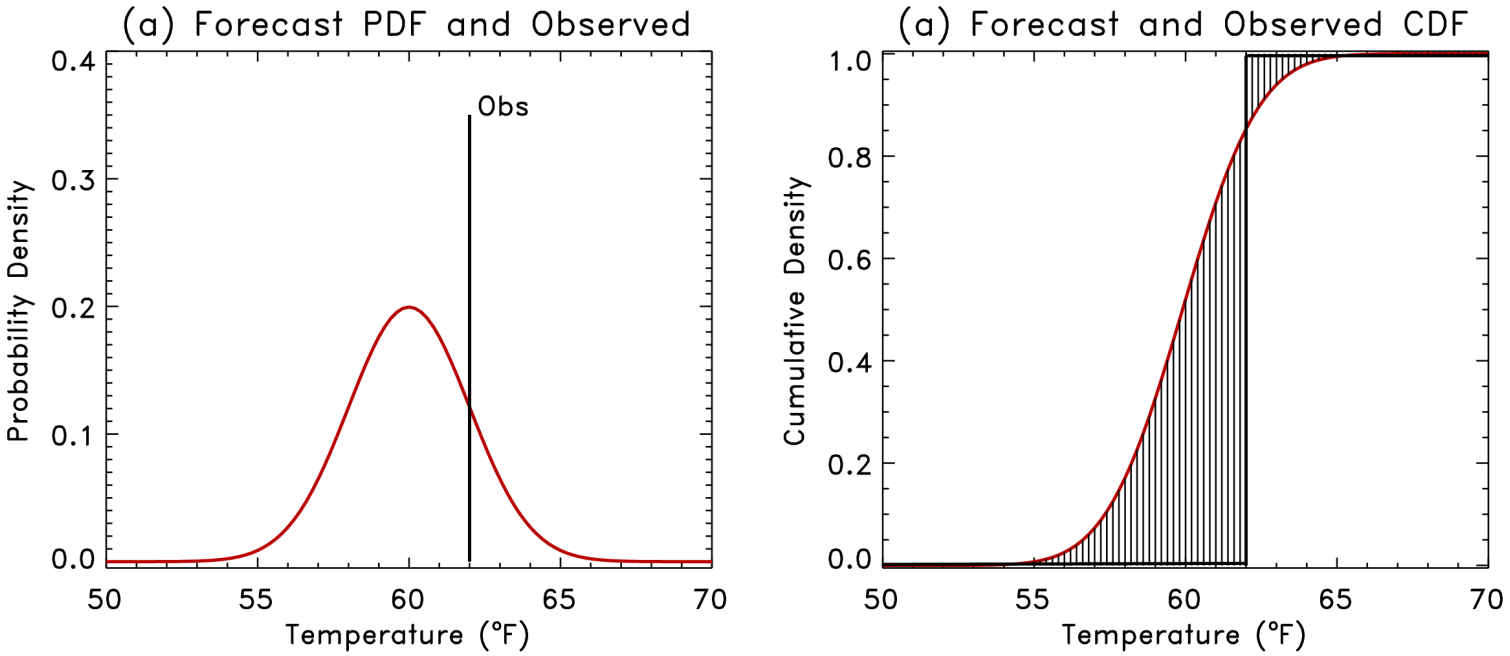
Warnings about rank histograms

- Assume all samples come from the same climotology!
- 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.

- Introduce pdf -> cdf High, low, no variance (event)
- Area of wide, narrow
- Perfect forecast with bias …
- Aggregate
- Relates to Brier score –

Ranked Probability Score – want to minimize area



- For a forecast of a binary event, the RPS score is equivalent to the Brier score.

MET tutorial

June 27 – 28

NCAR Foothills Lab

Registration and Information:

http://www.dtcenter.org/events/workshops11/met_tutorial.php