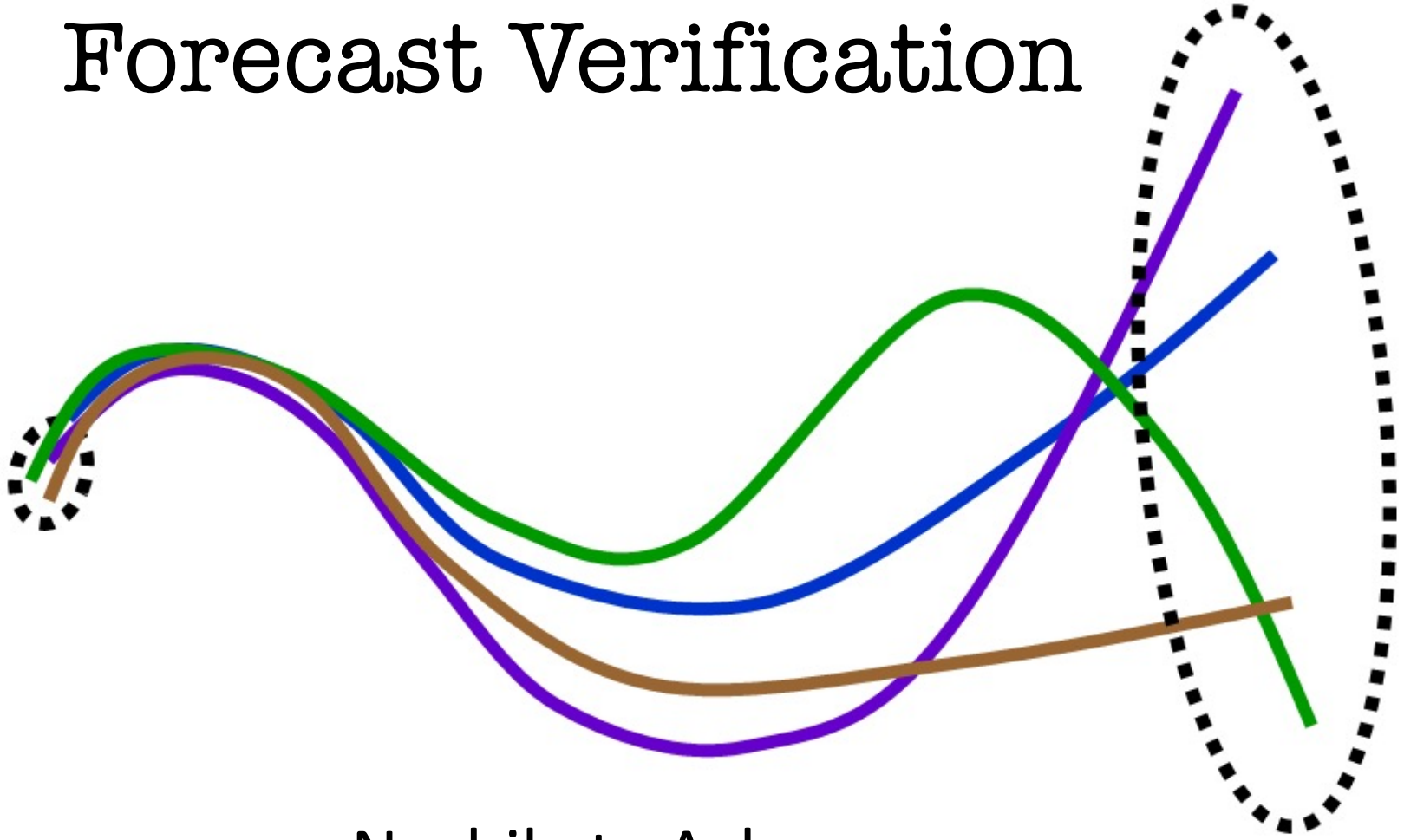
Ensemble & Probabilistic Forecast Verification



Nachiketa Acharya

npa5302@psu.edu

Center for Earth System Modeling, Analysis, and Data,

Department of Meteorology and Atmospheric Science, Pennsylvania State University, USA

PennState College of Earth and Mineral Sciences

Department of Meteorology and Atmospheric Science

ESMAD

Topics:

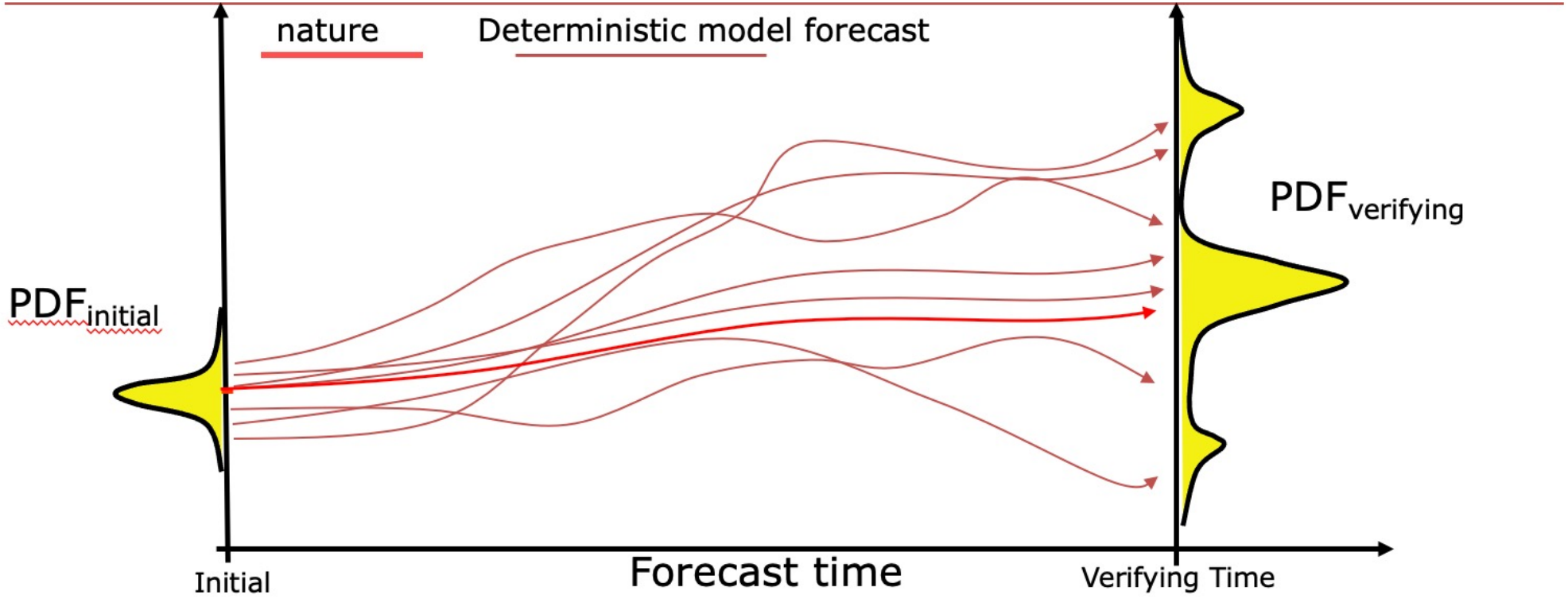
- Introduction to Ensemble and probabilistic forecast

- Ensemble prediction verification
  - Rank histogram
  - Ensemble spread relationship
  - Signal-to-noise ratio
  - Potential predictability

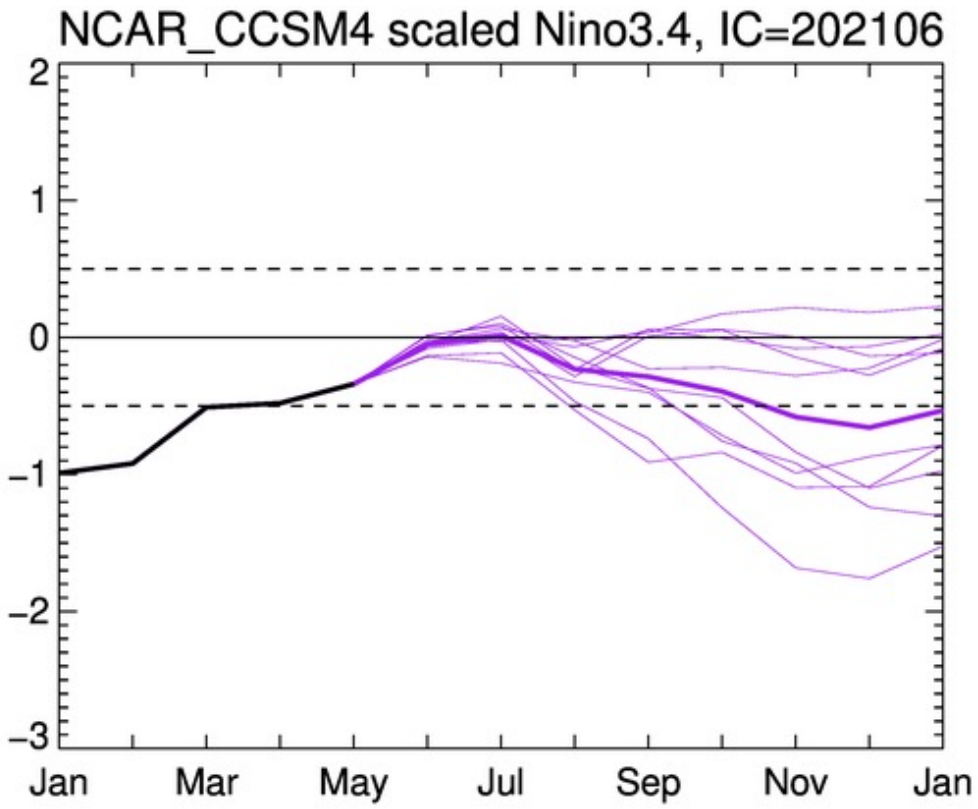
- Probabilistic prediction verification
  - BSS/RPSS/CRPSS including decompositions
  - logarithmic skill score
  - Reliability diagram
  - ROC/GROC

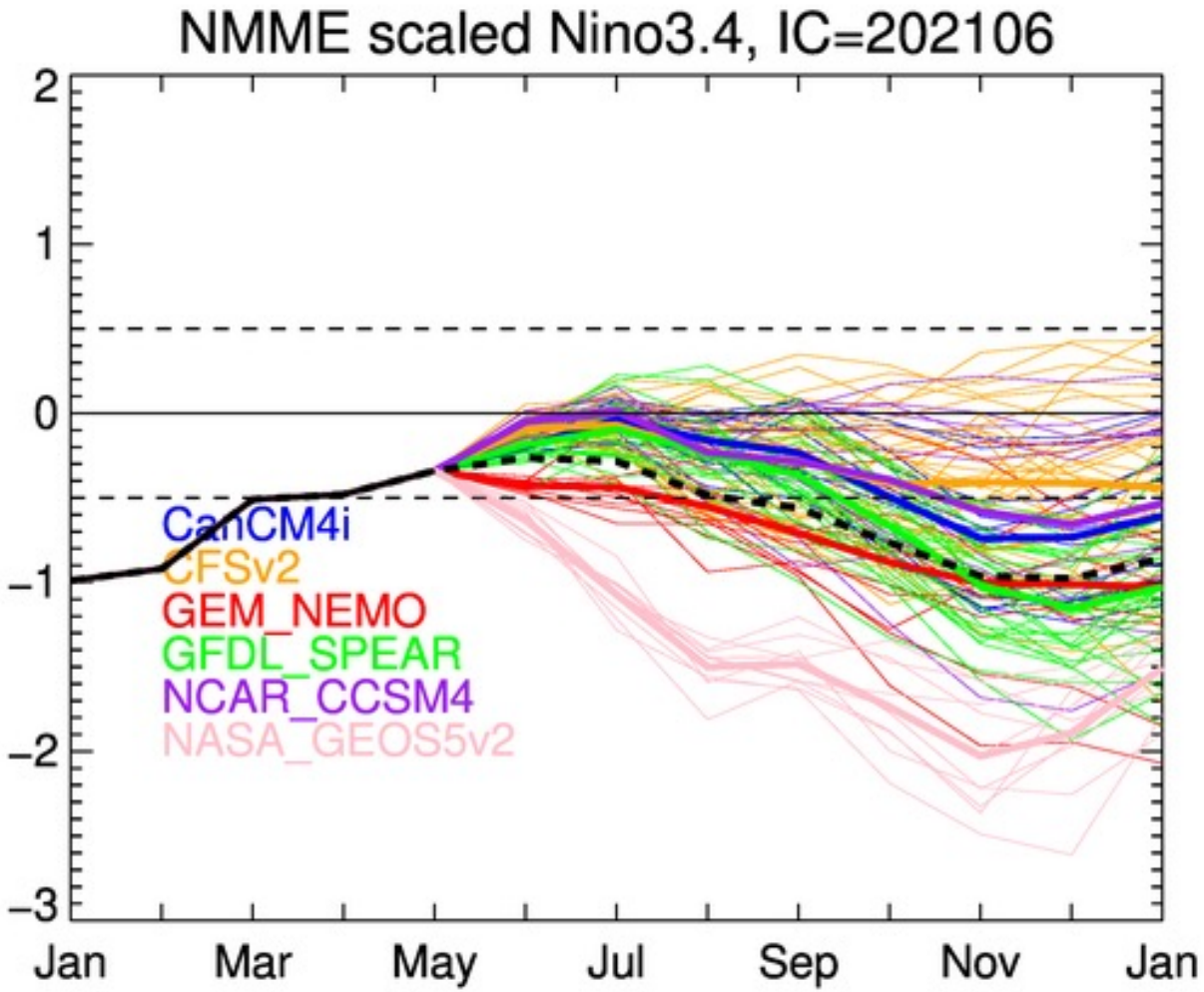
Ensemble forecast

- Instead of running just a single forecast, the dynamical model is run a number of times from slightly different starting conditions. The complete set of forecasts is referred to as the ensemble, and individual forecasts within it as ensemble members.
- Ensemble forecast systems are designed so that each member should be equally likely. The initial differences between the ensemble members are small, and consistent with uncertainties in the observations. But when we look several days ahead the forecasts can be quite different.

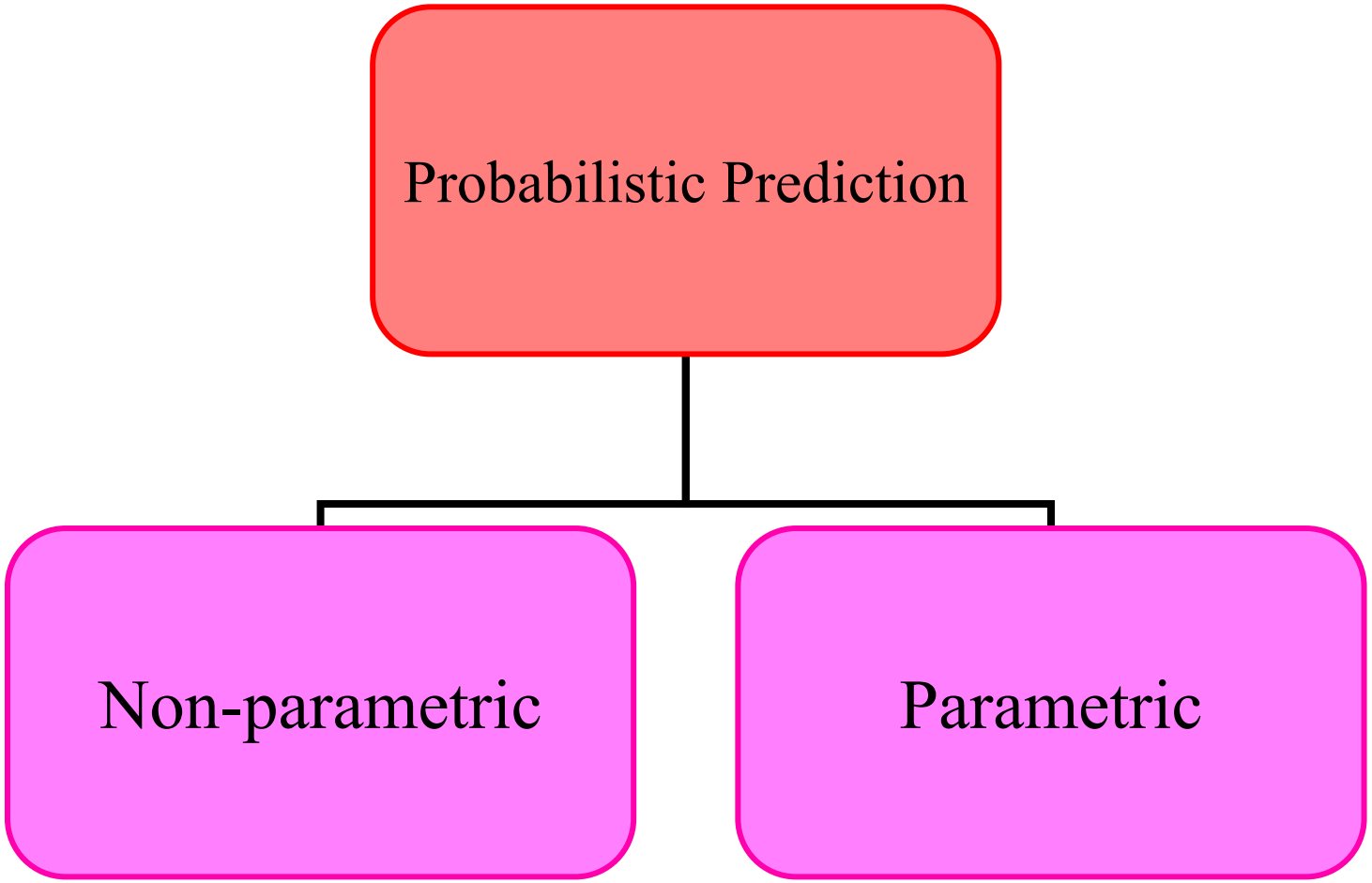


Visualization of Ensemble Forecasts





Probability Forecast based on Ensembles



Concept of probabilistic forecast based on parametric method

Tercile Category

- Below Normal
- Near Normal
- Above Normal

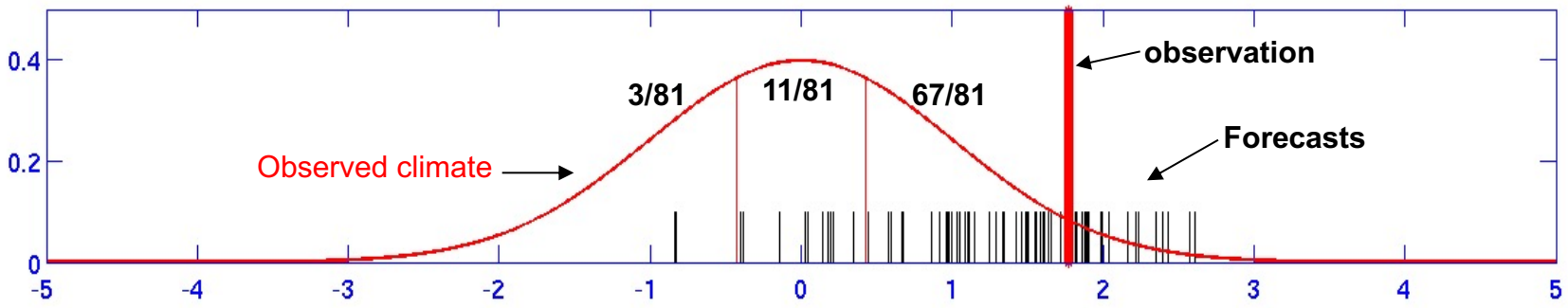
- **Below the 33th percentile is regarded as below-normal (BN) category,**
- **Above the 67th percentile is regarded as above-normal (AN) category,**
- **Between 33th and 67th percentiles is regarded as near-normal (NN) category.**

$$P(\Omega \mid GCM_i) = \frac{m_i}{M}$$

where, Ω = any of the tercile category (BN or NN or AN).
GCM_i = ith GCM.
m_i = ith ensemble member belongs to Ω category.
M = total number of ensemble members of GCM

Quantitative description

Example: An 81-members ensemble forecast is issued to predict the likelihood that a variable will verify in the upper tercile of the historical distribution (variable's climate). Let's call this event x. Looking at the diagram below we simply count the number of members falling in that tercile.



$$P(x) \approx \frac{n_x}{n_t} = \frac{67}{81}$$

Concept of probabilistic forecast based on parametric method

$$X = \beta + \varepsilon$$

where

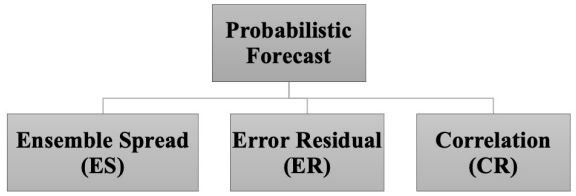
X is the forecast to be given,

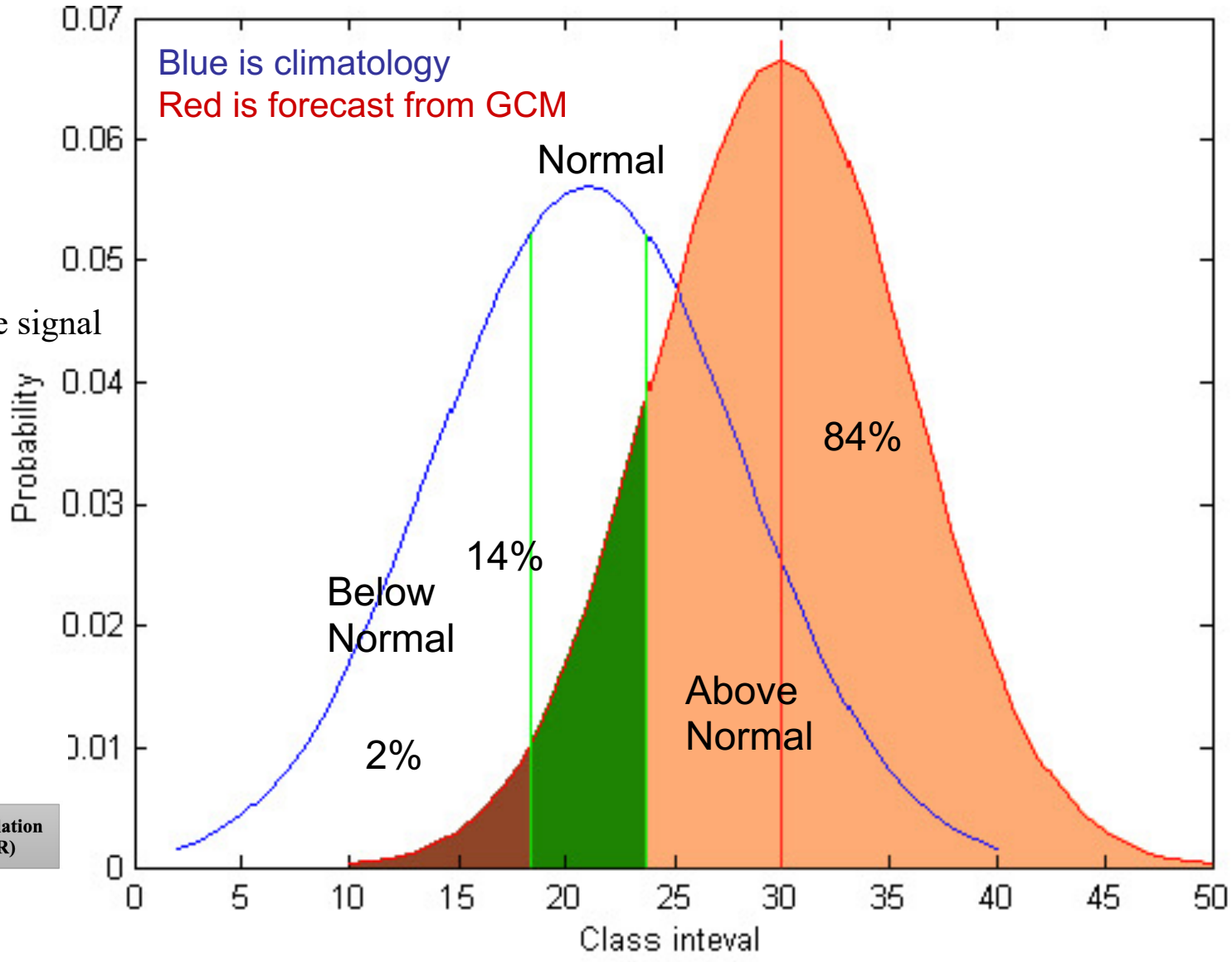
β is the potentially predictable signal

ε is the error part

β = Mean of member

σ_ε = ????





Ensemble prediction verification

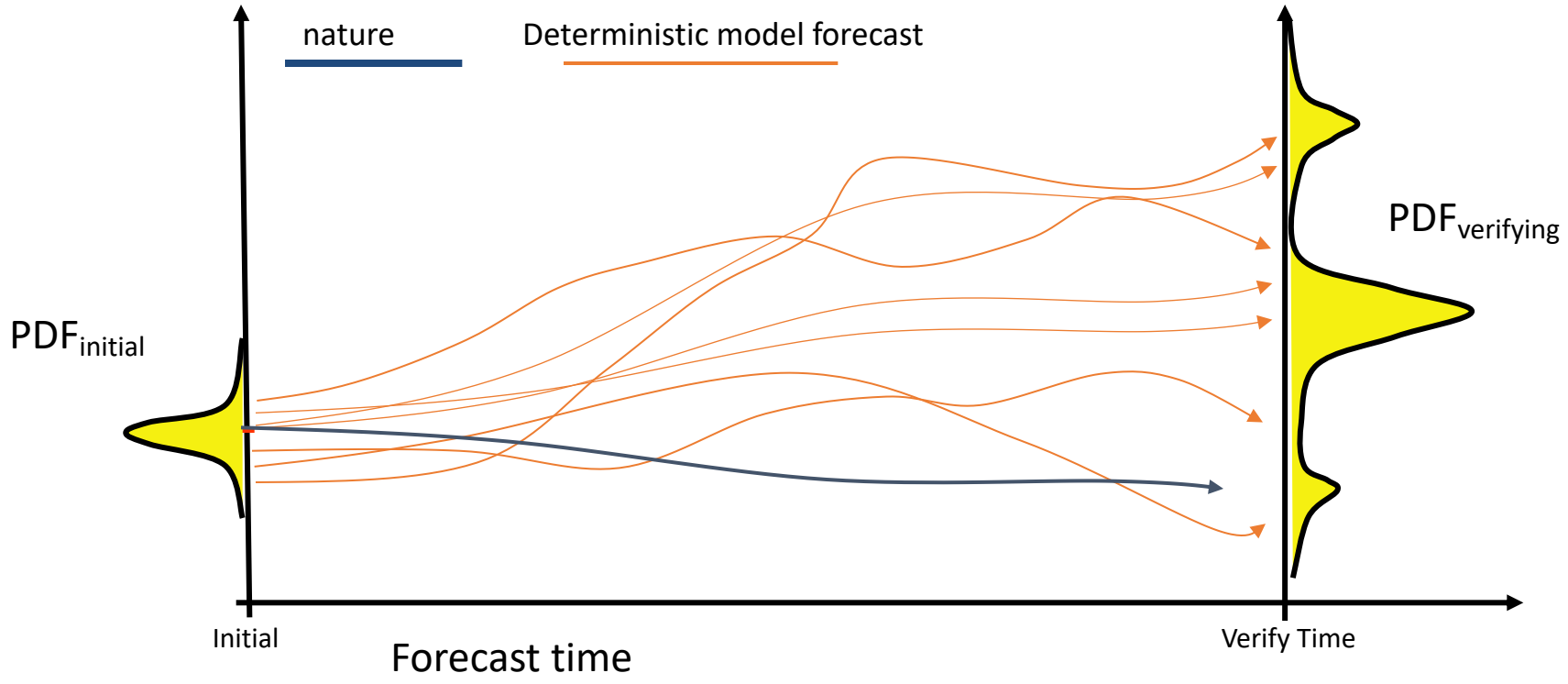
Do the observations statistically belong to the distributions of the forecast ensembles?

What is the utility of ensemble prediction systems?

Where the predictability of ensemble system comes?

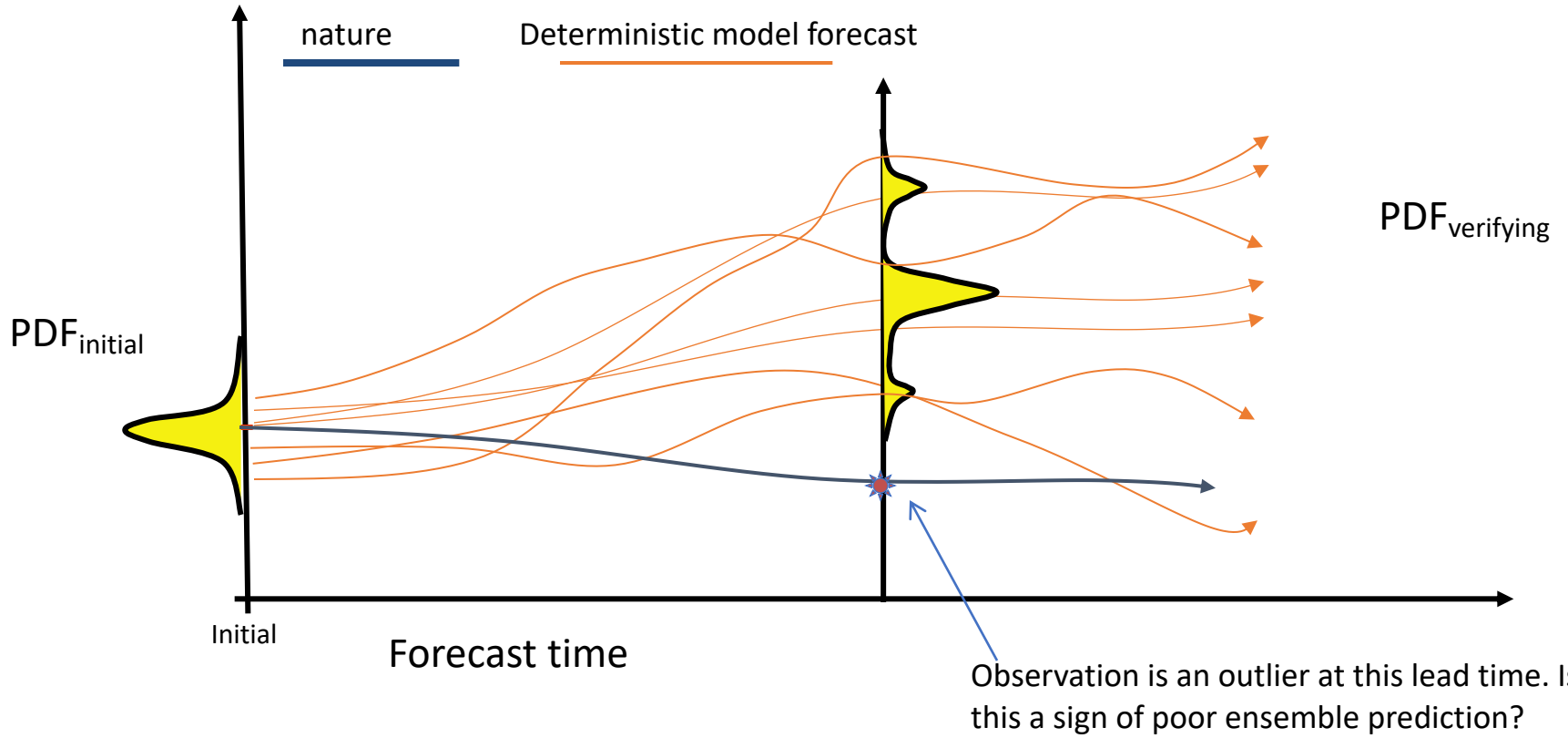
Rank Histogram (aka Talagrand Diagram)

Do the observations statistically belong to the distributions of the forecast ensembles?



Rank Histogram (aka Talagrand Diagram)

Do the observations statistically belong to the distributions of the forecast ensembles?



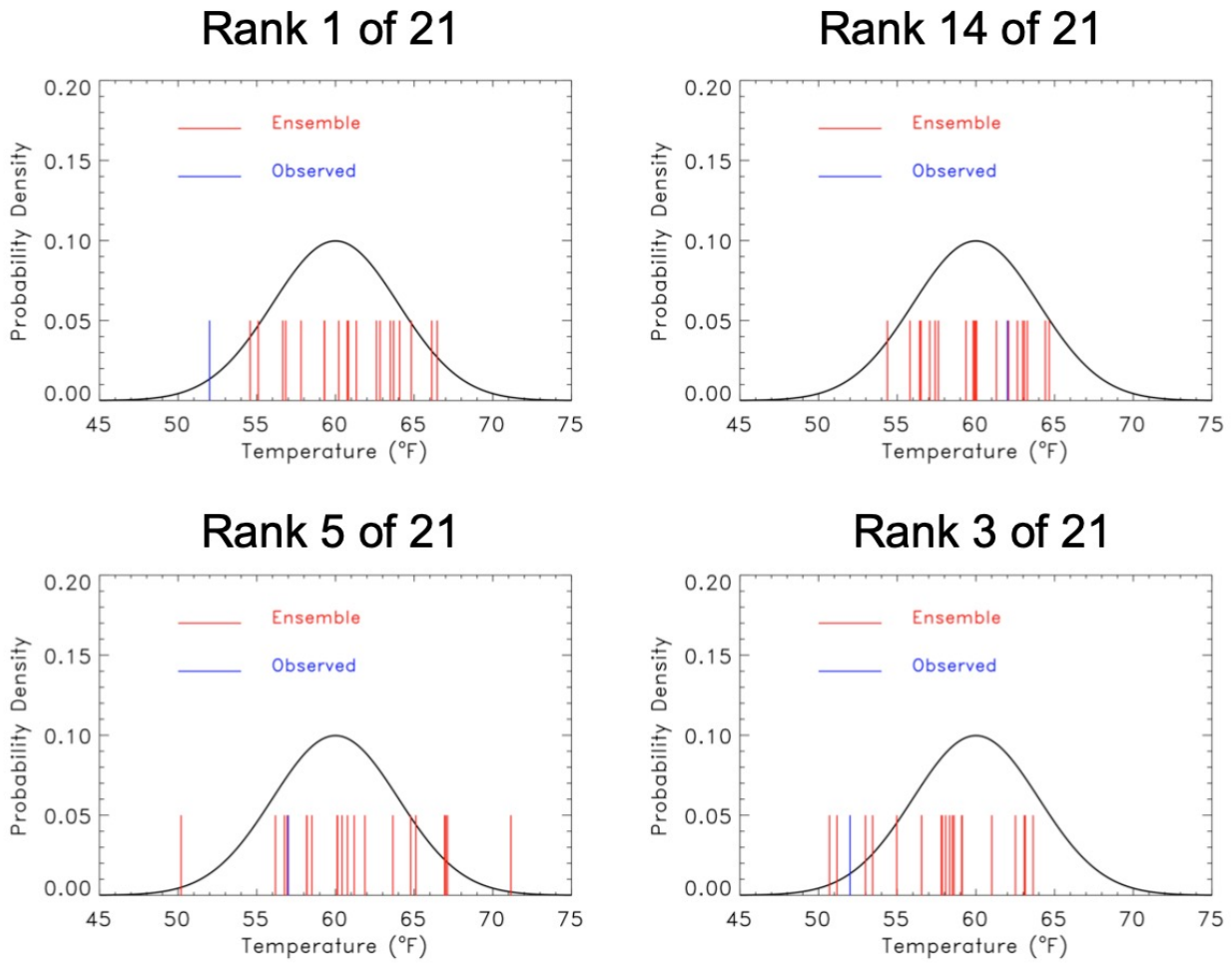
Rank Histogram (aka Talagrand Diagram)

- Assumption: Observation should in between ensemble spread of an ensemble system. This implies that observation should behave like a member of the model.

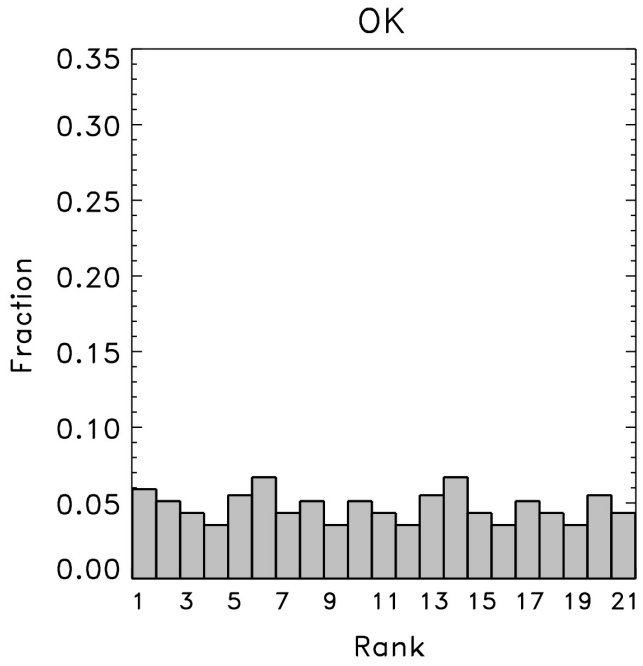
- Rank Histogram or Talagrand diagrams is a diagnostic tool to evaluate the spread of the ensemble members to measures consistency and reliability of a model (Hamil; 2001) .

- The main assumption behind is that the ensemble member forecasts are distributed so as to delineate ranges or "bins" of the predicted variable such that the probability of occurrence of the observation within each bin is equal.

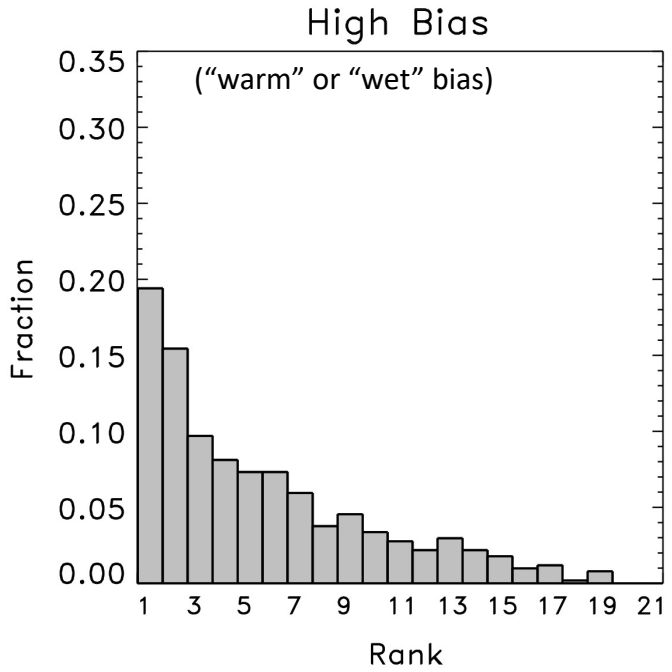
- For each specific forecast, the bins are determined by ranking the ensemble member forecasts from lowest to highest. The interval between each pair of ranked values forms a bin. If there are N ensemble members, then there will be (N+ 1) bin.



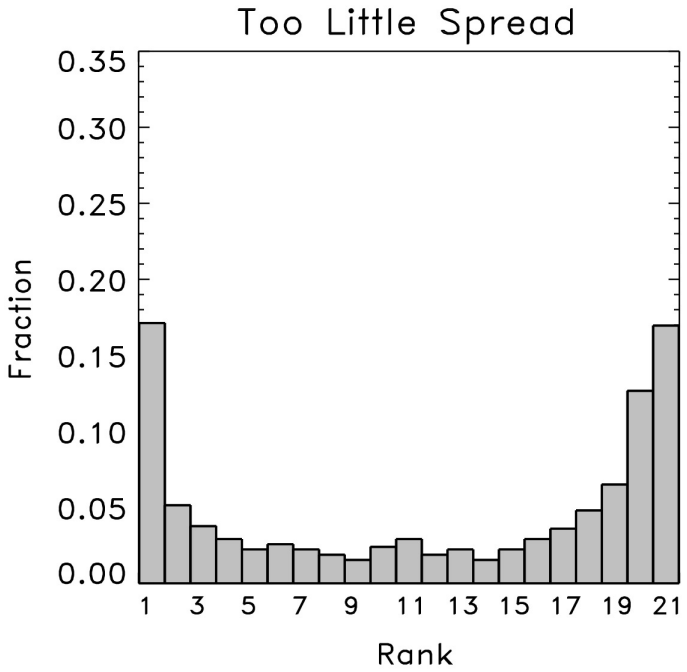
Rank Histograms



OBS is indistinguishable from any other ensemble member

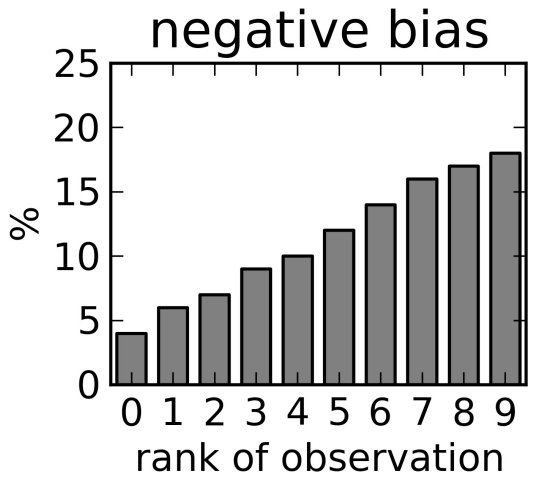


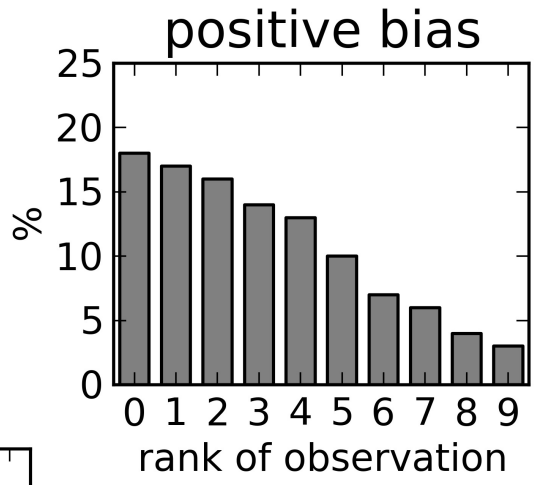
OBS is too often below the ensemble members (biased forecast)

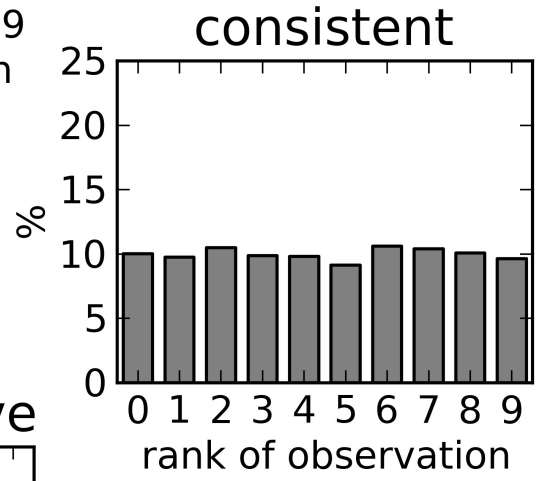


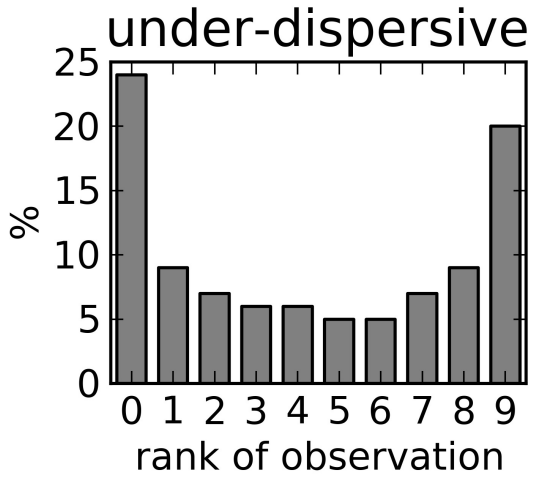
OBS is too often outside the ensemble spread

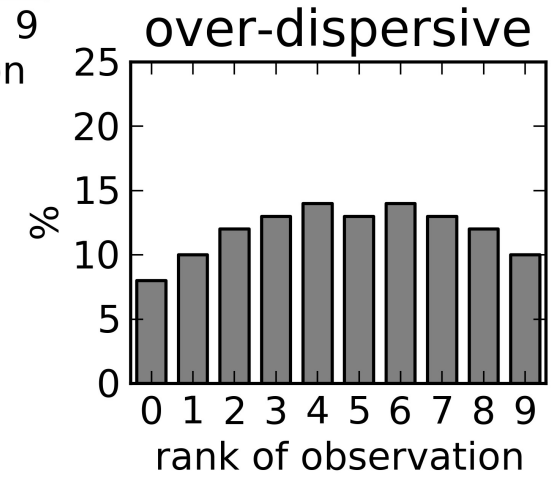
Rank Histograms





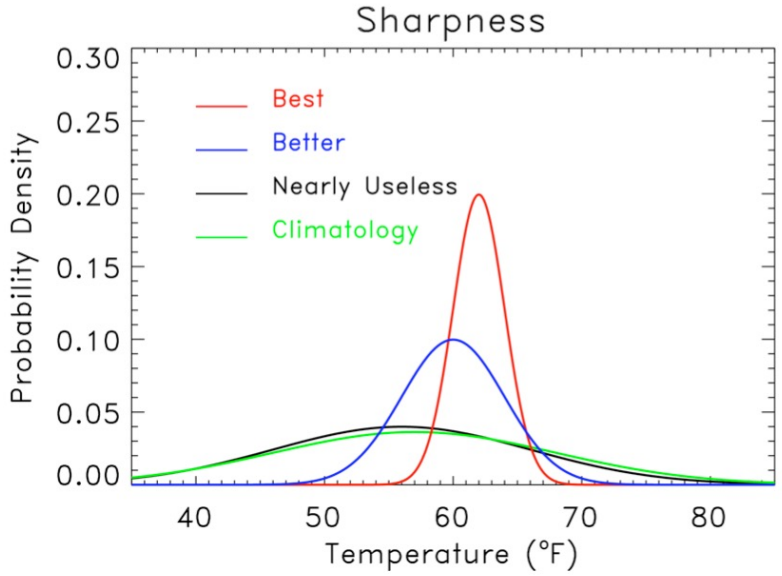






A uniform rank histogram is a necessary but not sufficient criterion for determining that the ensemble is reliable (see also: T. Hamill, 2001, MWR)

Rank histograms tell us about reliability - but what else is important?

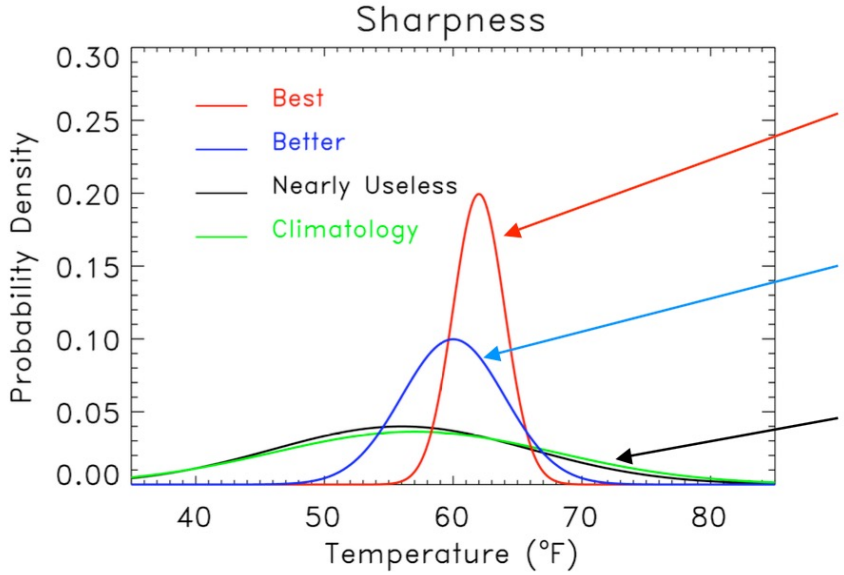


"Sharpness" measures the specificity of the probabilistic forecast. Given two reliable forecast systems, the one producing the sharper forecasts is preferable.

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

"Spread-skill" relationships are important, too.

Small-spread ensemble forecasts should have less ensemble-mean error than large-spread forecasts.



ensemble-mean error from a sample of this pdf on avg. should be low.

ensemble-mean error should be moderate on avg.

ensemble-mean error should be large on avg.

Spread-Skill Relationship: Measure of the utility of ensemble prediction systems

$$\text{spread} = \sqrt{\frac{\sum (\text{ensemble mean} - \text{ensemble forecast})^2}{n}}$$

$$\text{skill} = |\text{ensemble mean} - \text{observation}|$$

- As the ensemble members are interchangeable, the ensemble spread (standard deviation between ensemble members that measures the uncertainty of ensemble) should be representative of the uncertainty in the mean (the expected forecast error of ensemble-mean) which is popular as spread-skill relation (Whitaker and Loughe, 1998).

- Large (small) ensemble spread are required for less (more) predictable events, whereas less (more) predictable event is more (less) challenging to predict and should have broader (narrower) error range.

- Ensemble spread and absolute error of ensemble mean should be highly positively correlated on average while correlation less than 0.6 implies a poor representation of the uncertainty in the mean by the spread (Grimit and Mass, 2007).

Spread-Skill Relationship: Measure of the utility of ensemble prediction systems

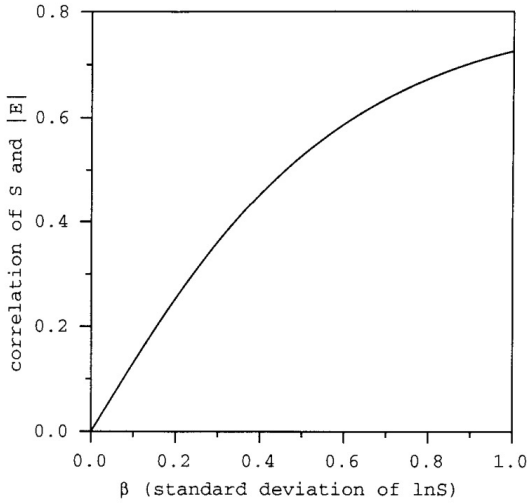
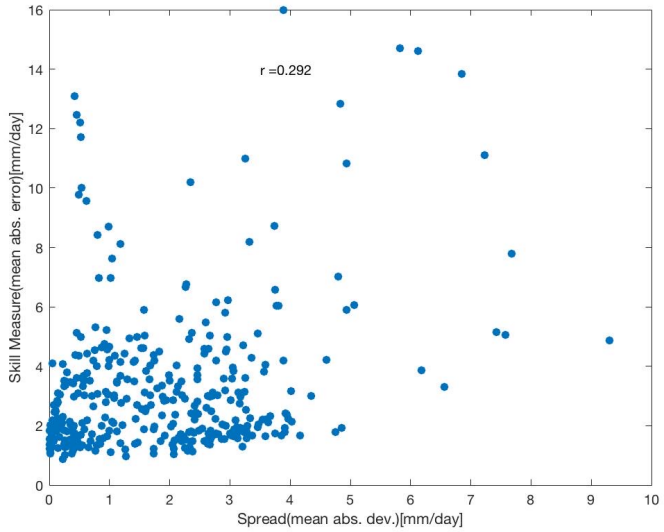
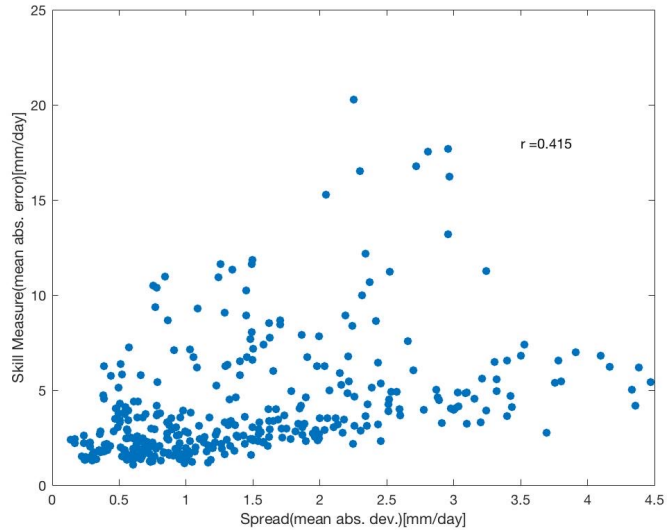
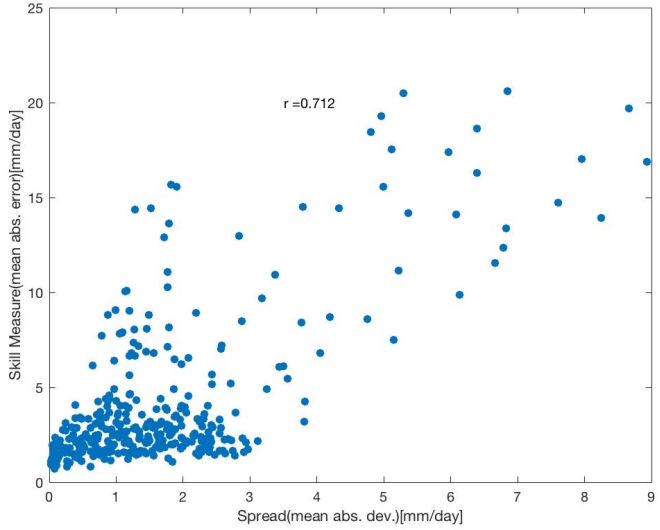


FIG. 1. Correlation of S and $|E|$ as a function of β, for the idealized statistical model given by (1) and $E = N(0, S)$.

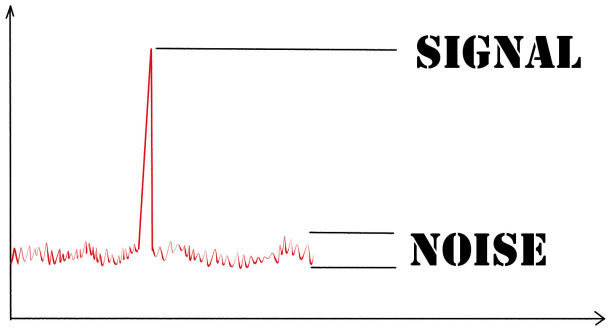
Whitaker and Loughe, 1998







Signal-to-Noise Ratio



- The predictability of ensemble system can be defined as the ratio of external and internal components.
- The ensemble mean of all the members is treated as the external component, and, the deviation of members from the ensemble mean is treated as the internal component

$$\text{SNR} = \frac{\text{Variance of ensemble mean}}{\text{Variance of intra-ensemble deviation}}$$

- The variance of the ensemble mean is a measure of the signal variability:

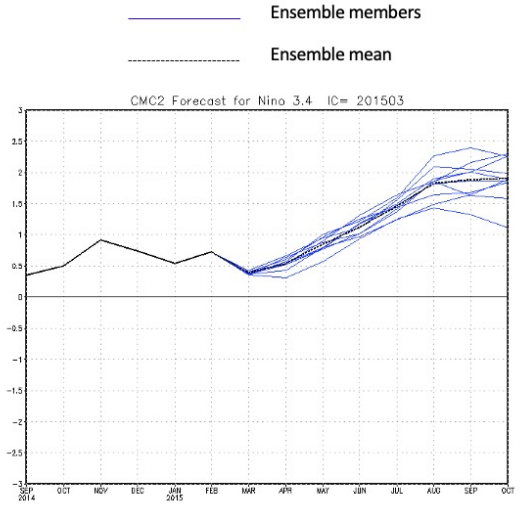
$$\overline{S}^2$$

- Variability of the ensemble spread is a measure of the noise. Define the variance of individual members wrt the ensemble mean (squared spread):

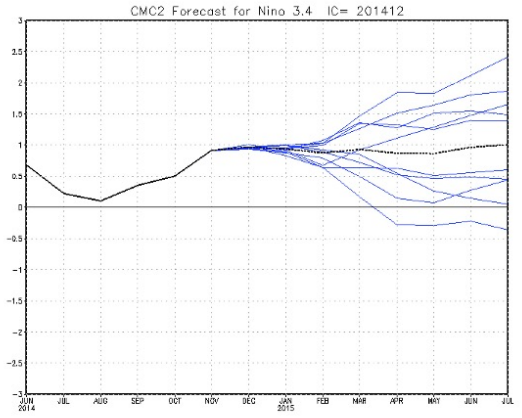
$$\overline{N}^2$$

- Then, the Signal to Noise Ratio for a particular initial condition is given as:

$$SNR = \sqrt{\frac{\overline{S}^2}{\overline{N}^2}}$$



Large SNR



Small SNR

Potential predictability or Signal to Total Ratio (STR)

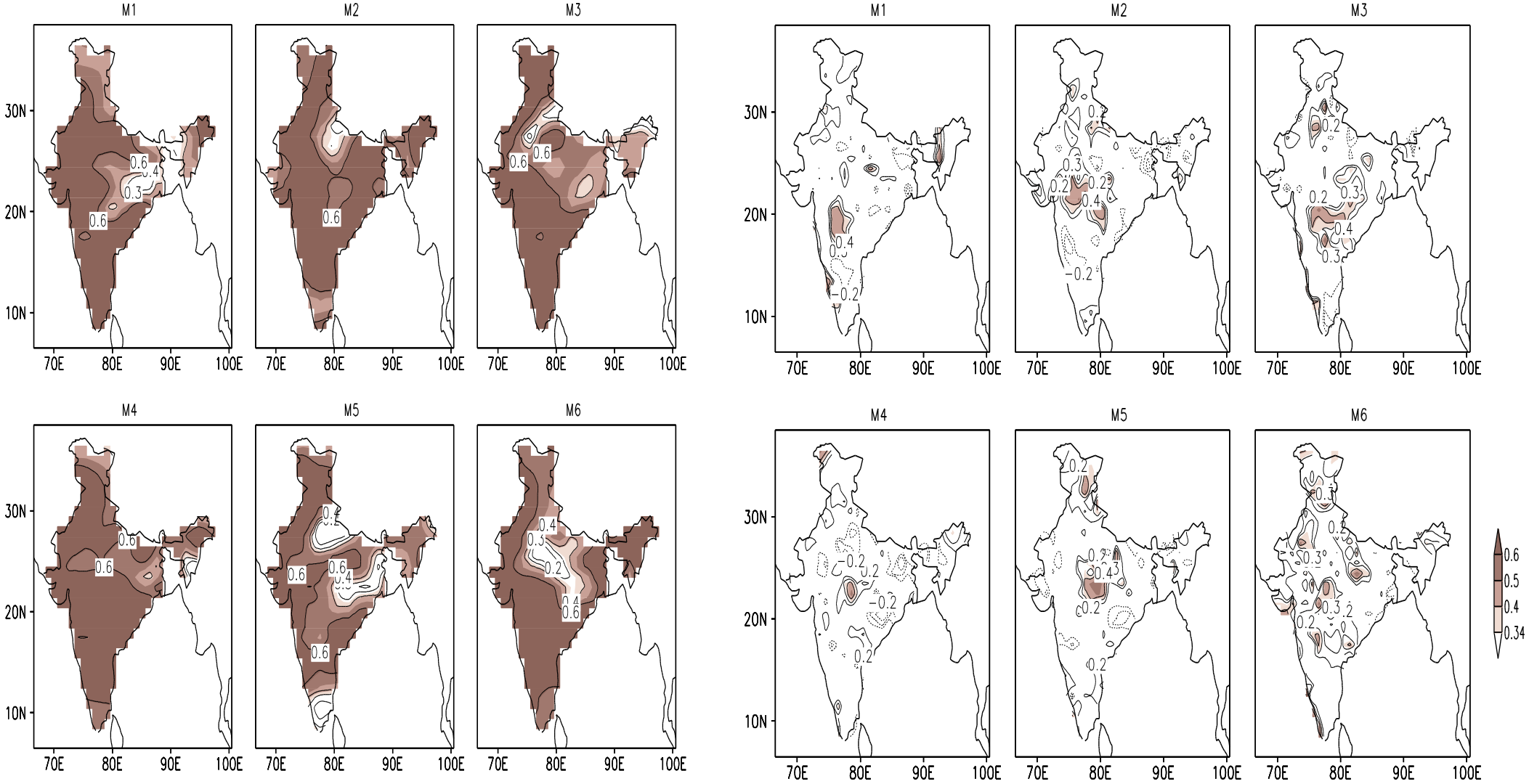
- The ratio of the variance of the potentially predictable signal ***(ensemble mean)*** to the total observed variances referred to as the potential predictability.

- For a perfect probability prediction ensemble, potential predictability or STR should high.

- In the perfect model framework, we assumed that an arbitrarily predicted ensemble member can be recognized as the “observation”.

- In other words, correlation ***r*** between the ensemble-mean and the observation should be same as correlation ***r*** between the ensemble-mean and any ensemble member.

Example: correlation of the ensemble members with the ensemble-mean (left hand) and the correlation of the observation with the ensemble mean (right hand)



Attributes of Probabilistic Forecast Quality

- **Accuracy:** Was the forecast close to what happened?
- **Reliability**: How well the *a priori* predicted probability forecast of an event coincides with the *a posteriori* observed frequency of the event
- **Resolution**: How much the forecasts differ from the climatological mean probabilities of the event, and the systems gets it right?
- **Sharpness**: How much do the forecasts differ from the climatological mean probabilities of the event?
- **Skill**: How much better are the forecasts compared to a reference prediction system (e.g., chance, climatology, persistence)?
- **Discrimination:** Are the forecasts differ when the outcome differs?

The Brier Score

- Mean square error of a probability forecast

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

where N is the number of realizations, p_i is the probability forecast of realization *i*. O_i is equal to 1 or 0 depending on whether the event (of realization *i*) occurred or not.

- Measures Forecast accuracy of Binary Events.
- Range: 0 to 1. Perfect=0
- Weighs larger errors more than smaller ones

Forecast Probability	Rain (Observed)
0.4229	No
0.0942	No
0.5985	Yes
0.4709	Yes
0.6959	No
0.6999	Yes
0.6385	Yes
0.0336	No
0.0688	Yes
0.3196	No

Forecast Probability	Rain (observed)
0.4229	0
0.0942	0
0.5985	1
0.4709	1
0.6959	0
0.6999	1
0.6385	1
0.0336	0
0.0688	1
0.3196	0

Components of the Brier Score

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

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.

- 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

Brier Score = Reliability – Resolution + Uncertainty

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

Brier Skill Score: Measures accuracy and skill respectively

Skill: Proportion of improvement of accuracy over the accuracy of a reference forecast (e.g., 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

- Cautions:
  - Cannot compare BS on different samples
  - BSS – Takes care of underlying climatology
  - BSS – Takes care of small samples

- Range: -Inf to 1; No skill beyond reference=0; Perfect score =1

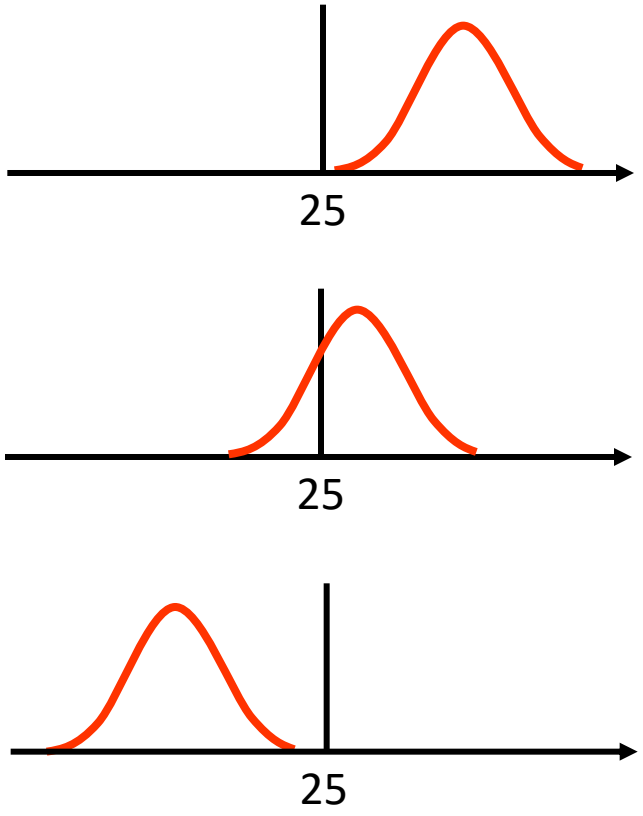
Reliability

- 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.
  - Example: Climatological forecast is reliable but does not provide any forecast information beyond climatology

- 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?

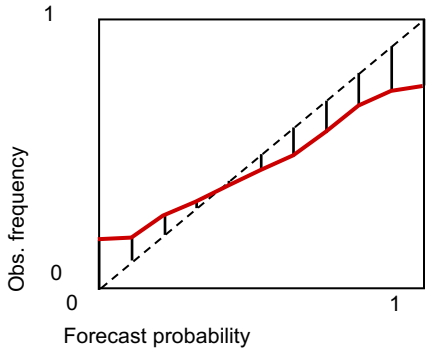


FC Prob.	# FC	OBS-Frequency (perfect model)	OBS-Frequency (imperfect model)
100%	8000	8000 (100%)	7200 (90%)
90%	5000	4500 (90%)	4000 (80%)
80%	4500	3600 (80%)	3000 (66%)
….	….	….	….
….	….	….	….
….	….	….	….
10%	5500	550 (10%)	800 (15%)
0%	7000	0 (0%)	700 (10%)

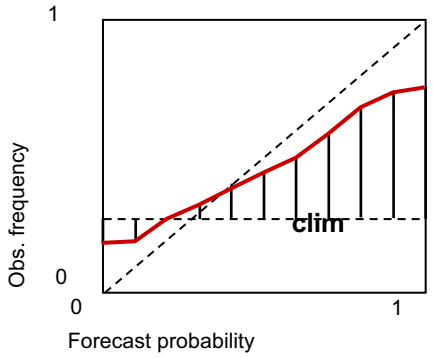
Reliability Diagram

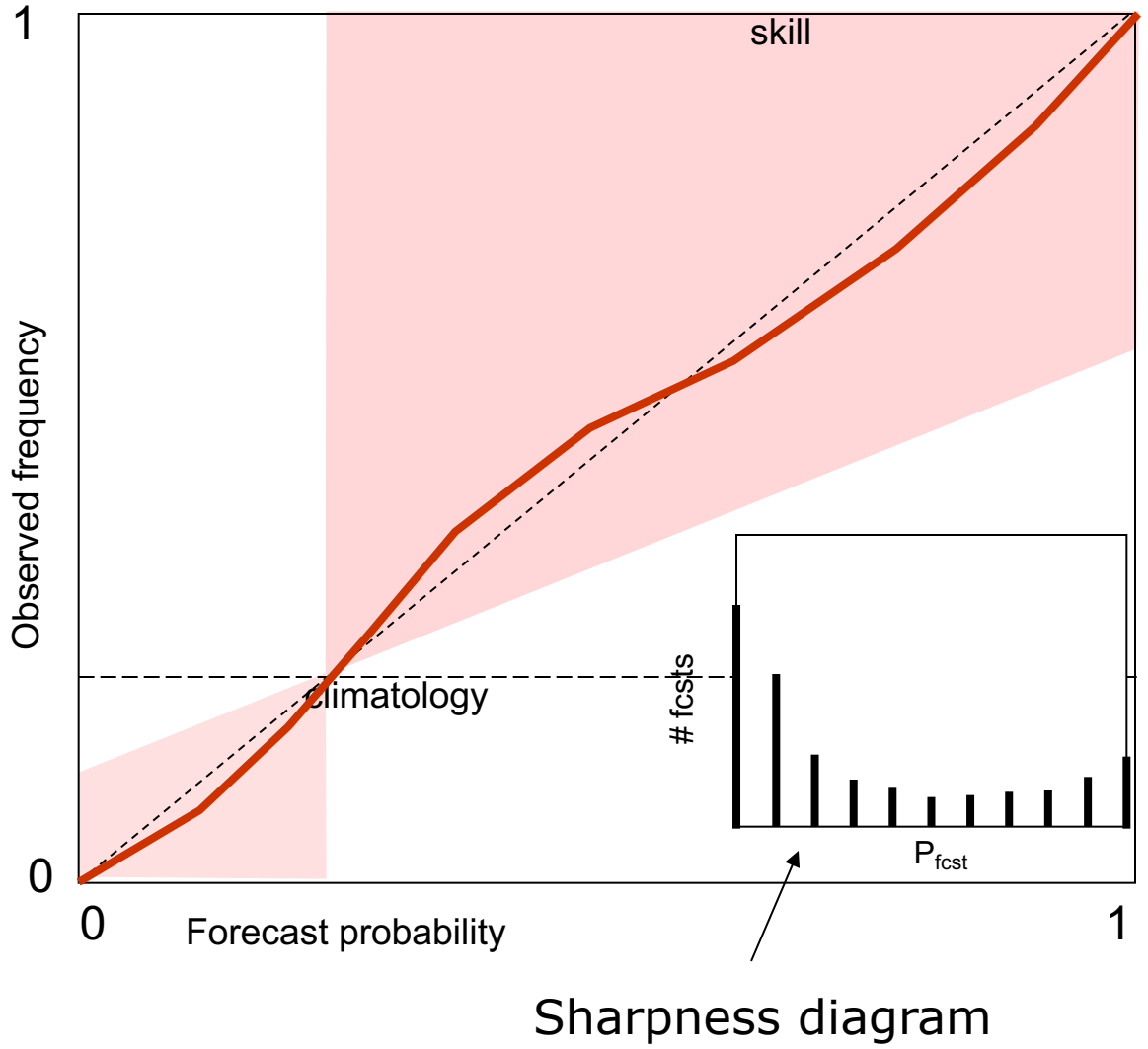
Reliability: How close to diagonal

(the lower the value the better)



Resolution: How far to horizontal (climatology) line





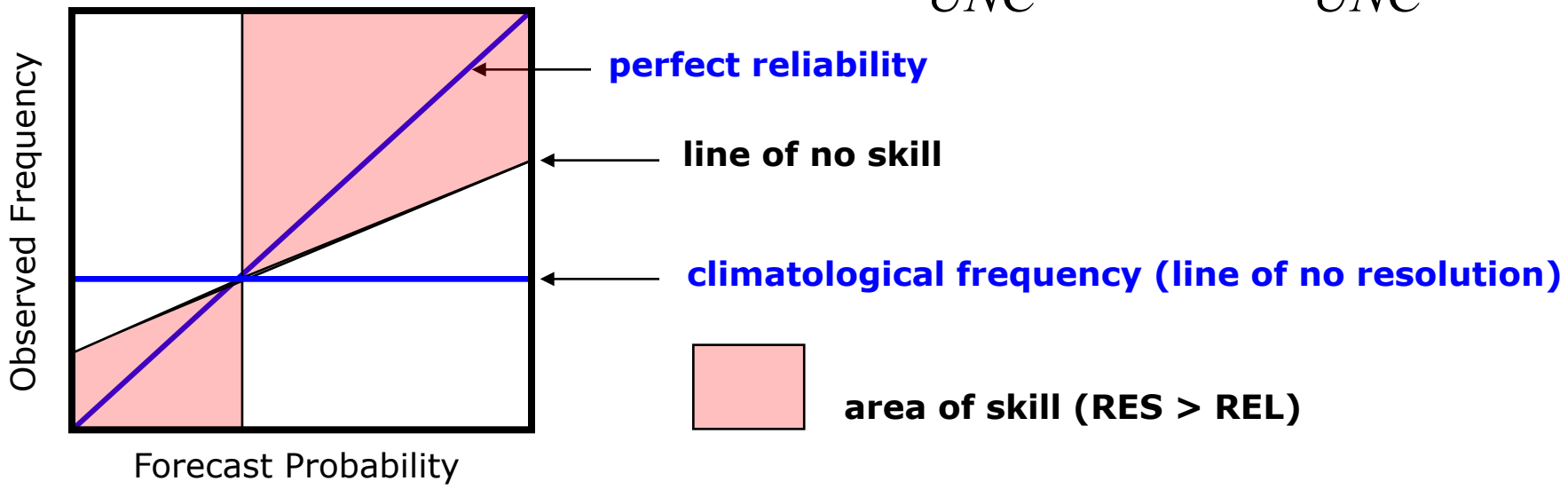
Sharpness diagram

Brier Skill Score & Reliability Diagram

- How to construct the area of positive skill?

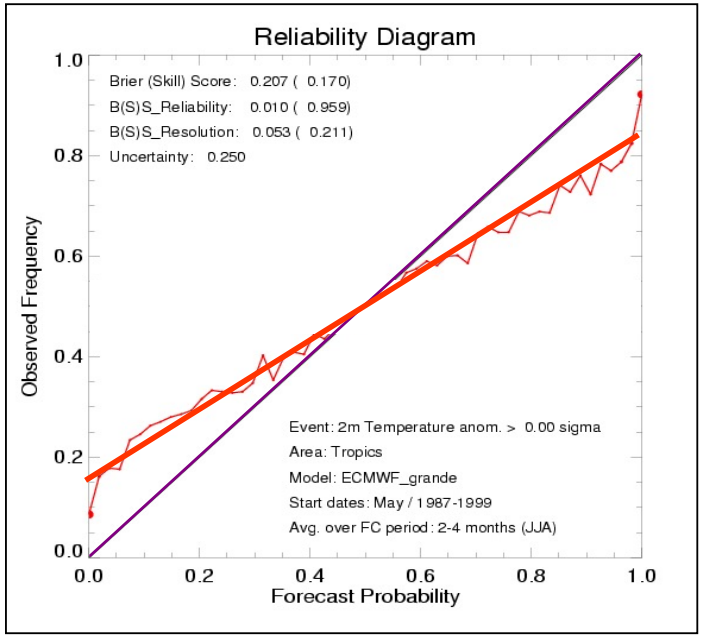
$$BSS = 1 - \frac{BS}{BS_c}$$

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

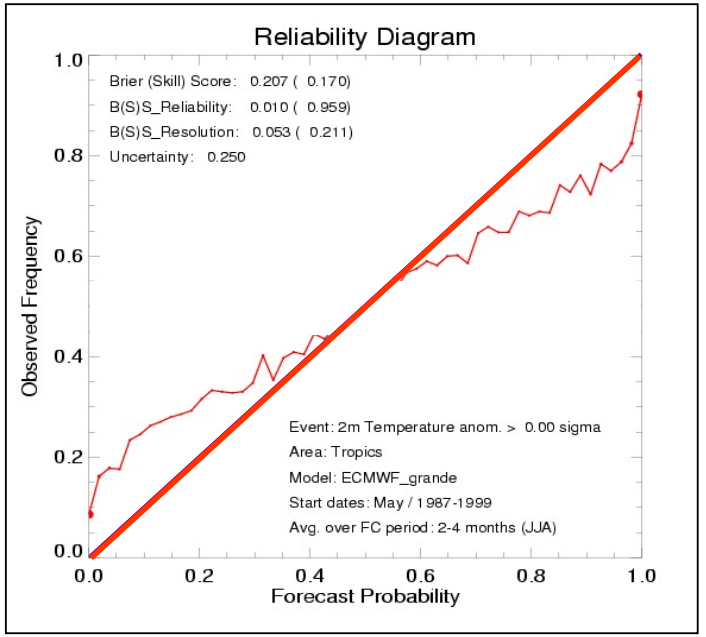


Examples of Reliability Diagram

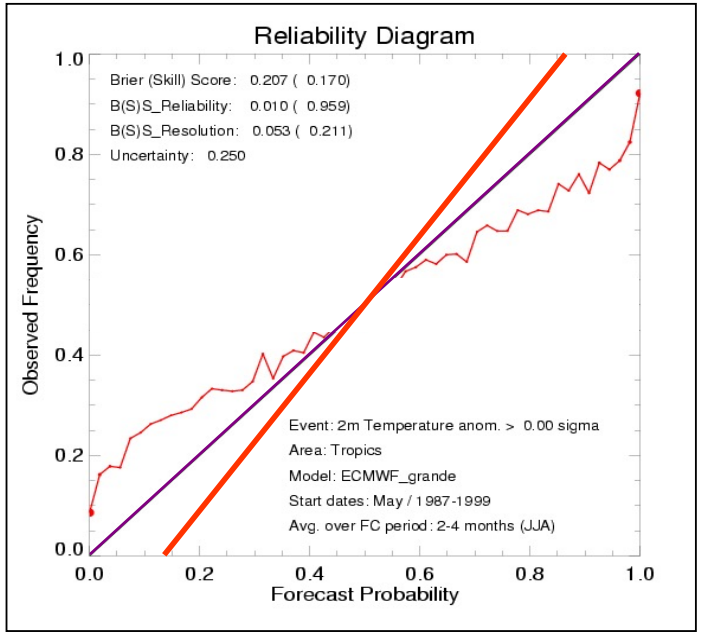
over-confident model



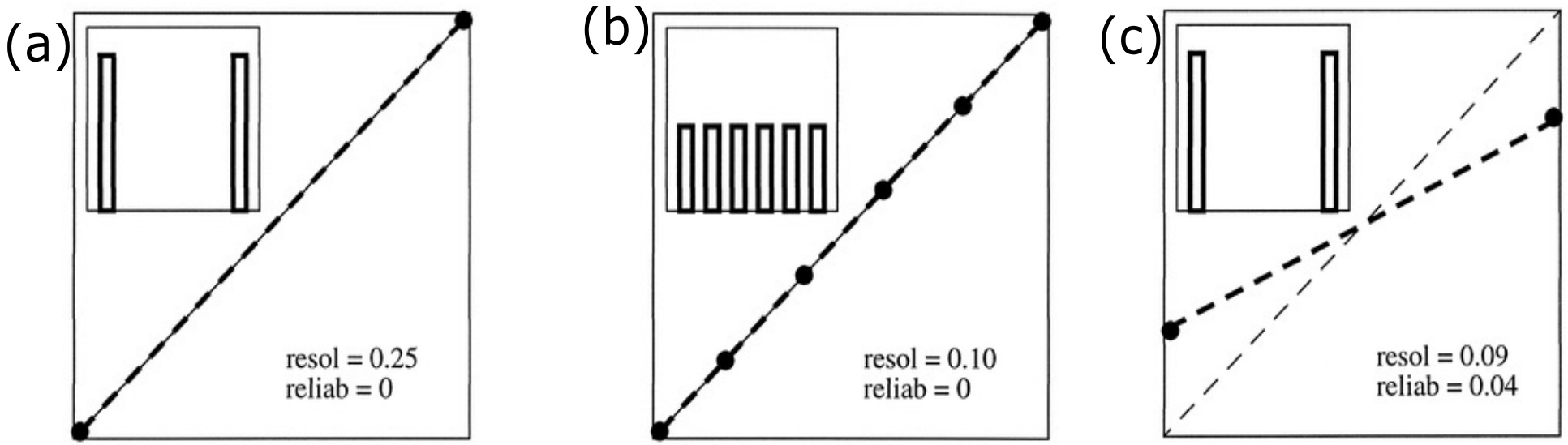
perfect model



under-confident model



Examples of Reliability Diagram



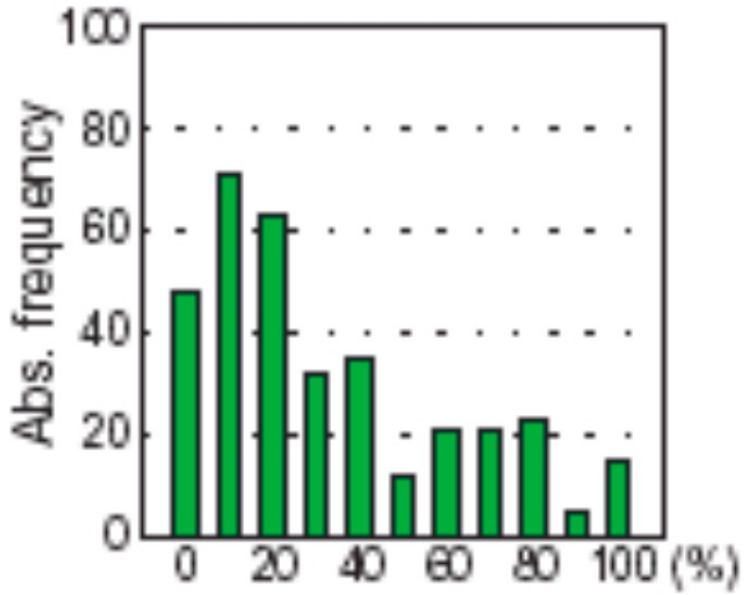
Typical reliability diagrams and sharpness histograms (showing the distribution of predicted probabilities). (a) Perfect resolution and reliability, perfect sharpness. (b) Perfect reliability but poor sharpness, lower resolutior than (a). (c) Perfect sharpness but poor reliability, lower resolution than (a).

Atger, 1999

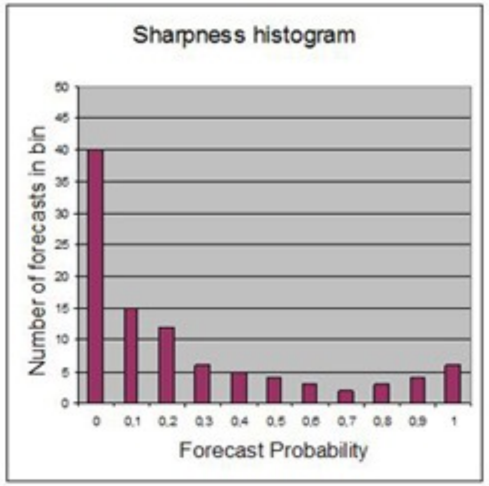
Sharpness

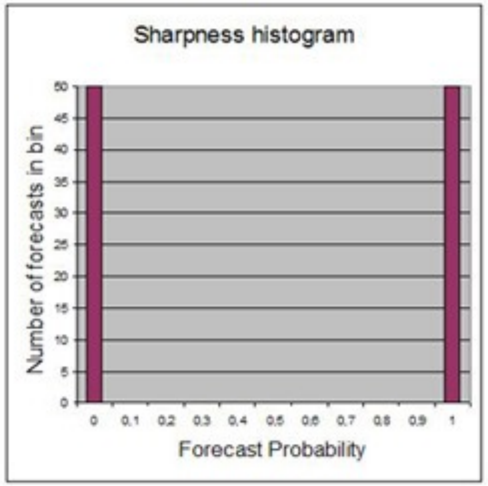
Refers to the spread of the probability distributions.

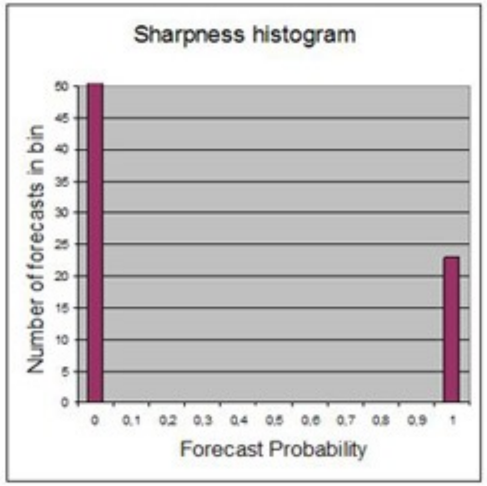
It is expressed as the capability of the system to forecast extreme values, or values close 0 or 1. The frequency of forecasts in each probability bin (shown in the histogram) shows the sharpness of the forecast.

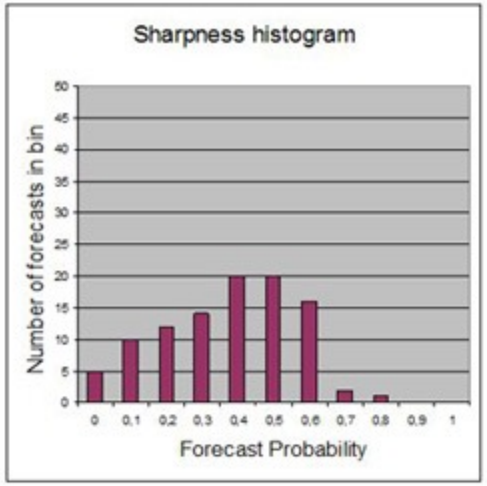


Sharpness Histogram Example









Reliability Diagrams - Comments

- Graphical representation of Brier score components
- Measures “reliability”, “resolution” and “sharpness”
- Sometimes called “attributes” diagram.
- Large sample size required to partition (bin) into subsamples conditional on forecast probability

Verification of Probabilistic Categorical Forecasts:
The **Ranked Probability Skill Score (RPSS)**

Epstein (1969), J. Appl. Meteor.

RPSS measures cumulative squared error between categorical forecast probabilities and the observed categorical probabilities relative to a reference (or standard baseline) forecast.

The observed categorical probabilities are 100% in the observed category, and 0% in all other categories.

$$RPS = \sum_{cat=1}^{Ncat} (Pcum_{F(cat)} - Pcum_{O(cat)})^2$$

Where Ncat = 3 for tercile forecasts. The “cum” implies that the summation is done for cat 1, then cat 1 and 2, then cat 1 and 2 and 3.

$$RPS = \sum_{cat=1}^{Ncat} (Pcum_{F(cat)} - Pcum_{O(cat)})^2$$

The higher the RPS, the poorer the forecast. RPS=0 means that the probability was 100% given to the category that was observed.

The RPSS is the RPS for the forecast compared to the RPS for a **reference forecast** such as one that gives climatological probabilities.

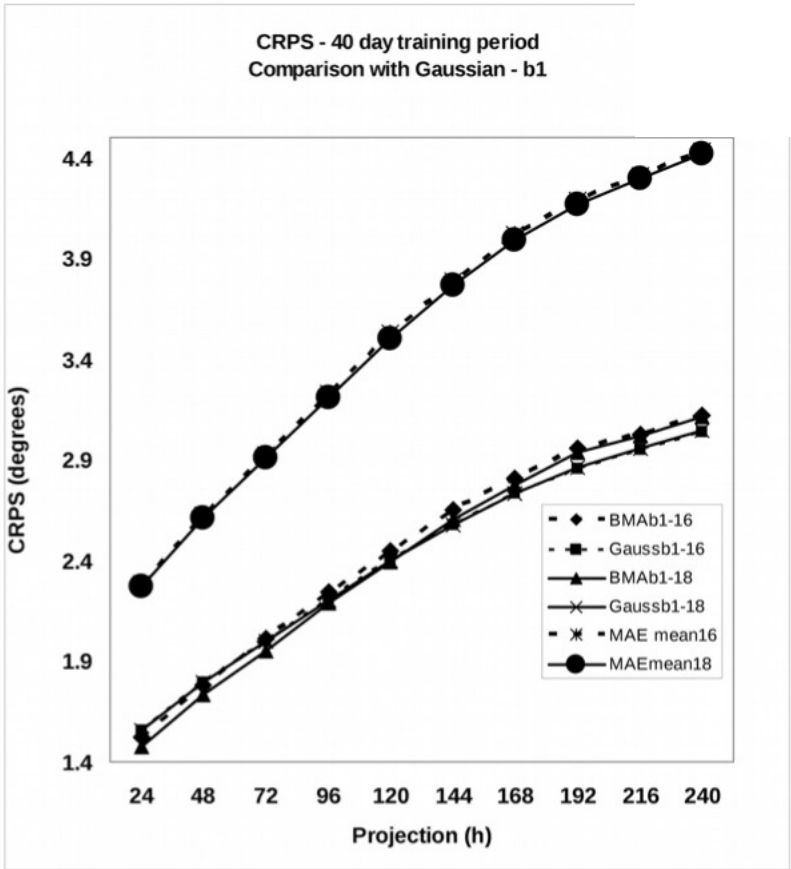
$$RPSS = 1 - \frac{RPS_{forecast}}{RPS_{reference}}$$

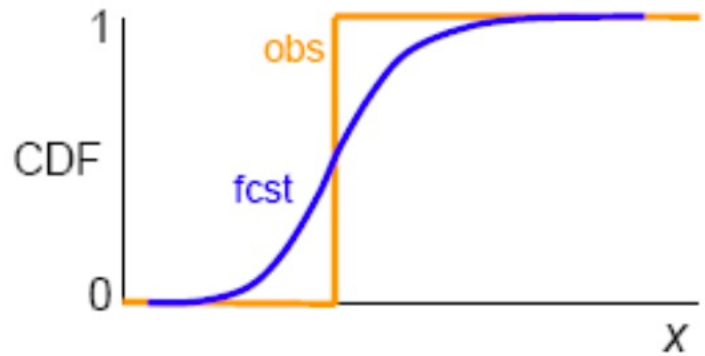
RPSS > 0 when RPS for actual forecast is smaller than RPS for the reference forecast.

Continuous Rank Probability Score

- Let $F_i^f(x)$ be the forecast probability CDF for the *i*th forecast case.
- Let $F_i^o(x)$ be the observed probability CDF (Heaviside function).

$$CRPS(forecast) = \frac{1}{ncases} \sum_{i=1}^{ncases} \int_{x=-\infty}^{x=-\infty} \left(F_i^f(x) - F_i^o(x) \right)^2 dx$$

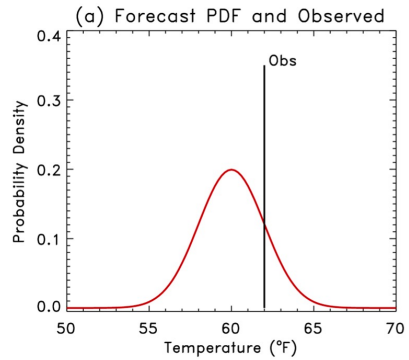


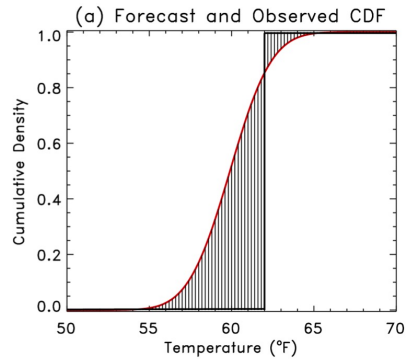


-difference between observation and forecast, expressed as cdfs

-defaults to MAE for deterministic fcst

-flexible, can accommodate uncertain obs





Continuous ranked probability *skill* score (CRPSS)

Like the Brier score, it's common to convert this to a skill score by normalizing by the skill of climatology, or some other reference.

$$CRPSS = \frac{\overline{CRPS}(forecast) - \overline{CRPS}(climo)}{\overline{CRPS}(perfect) - \overline{CRPS}(climo)}$$

Logarithmic skill score (LSS)

- The logarithmic score (LS) is an alternative to the RPS and for an individual forecast is the simply the logarithm of the forecast probability of the observed category.
- The LS does not depend on the forecast probabilities of the other categories, and this property means that the LS is a local score, unlike RPS.
- If the observations falls in the i-th category, then the LS of the forecast is $\log(p_i)$.
- logarithmic skill score (LSS) defined as $\mathrm{LSS} = \overline{\mathrm{LS}} - \overline{\mathrm{LS}}_{\mathrm{ref}}\,,$
- where LS and $\mathrm{LS}_{\mathrm{ref}}$ are the LS of the forecast being evaluated and a reference forecast, respectively, averaged over multiple forecasts.

- 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) SNR, STR

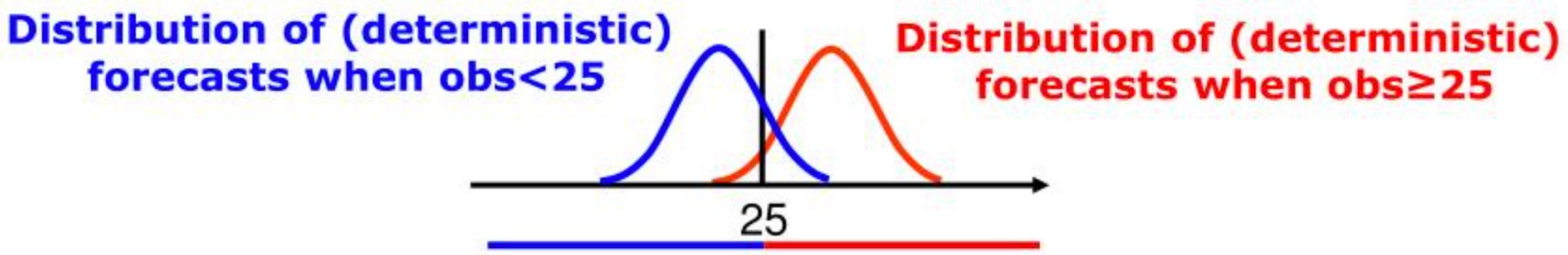
Reliability Diagram

- **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 even, and the systems gets it right?

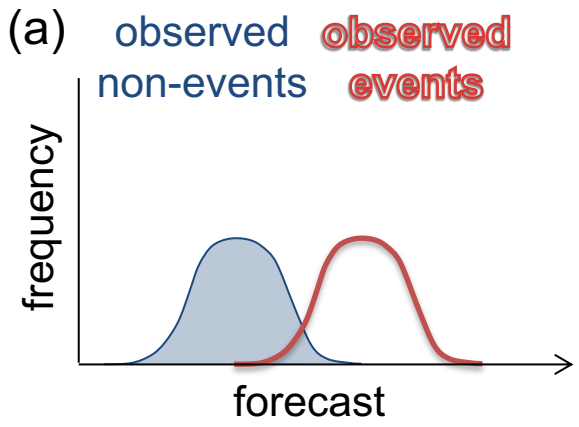
- **Skill**: Are the forecasts better than my reference system (chance, climatology, persistence,…)? BSS, RPSS, LSS

Discrimination

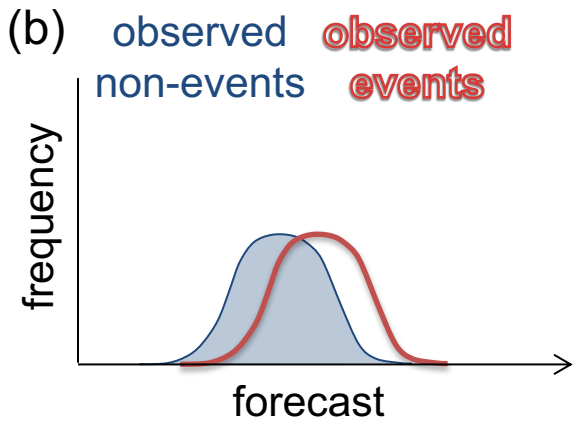
- Until now, we looked at the question:
  - If the forecast system predicts *x*, what is the observation *y*?
- When we are interested in the ability of a forecast system to discriminate between events and non-events, we investigate the question:
  - If the event *y* occurred, what was the forecast *x*?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



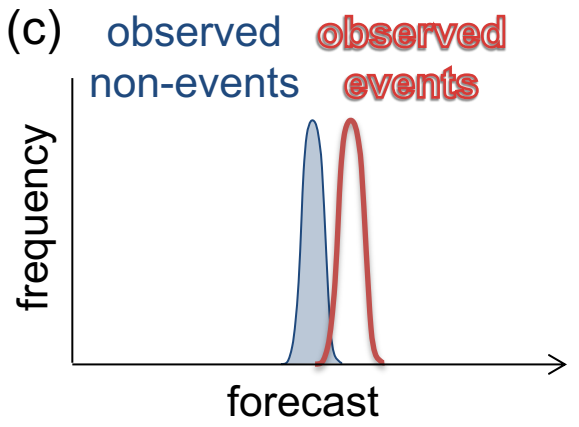
Discrimination



Good discrimination



Poor discrimination

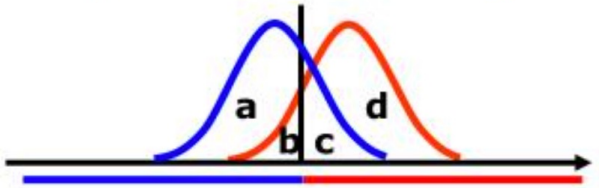


Good discrimination

Discrimination

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

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

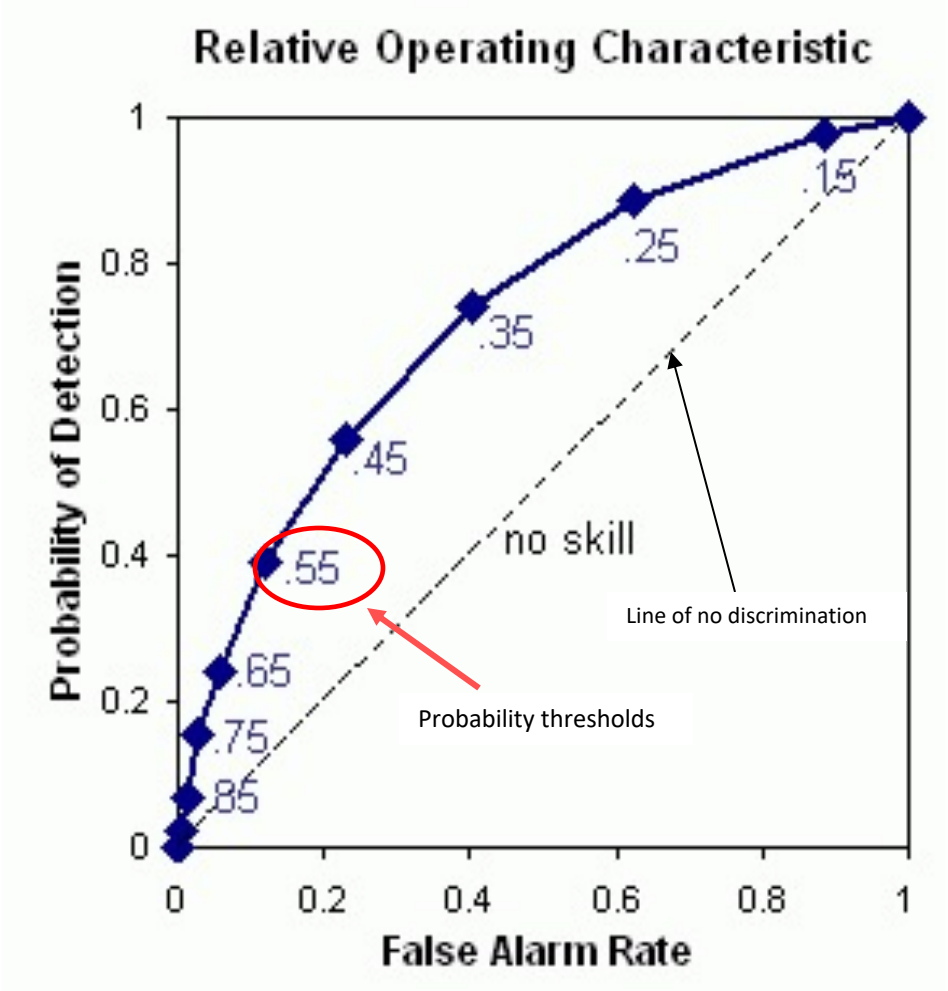


- Event Probability: $s = (a+c) / n$
- Probability of a Forecast of occurrence: $r = (a+b) / n$
- Frequency Bias: $B = (a+b) / (a+c)$
- Proportion Correct: $PC = (a+d) / n$

- Hit Rate: $H = a / (a+c)$
- False Alarm Rate: $F = b / (b+d)$
- False Alarm Ratio: $FAR = b / (a+b)$

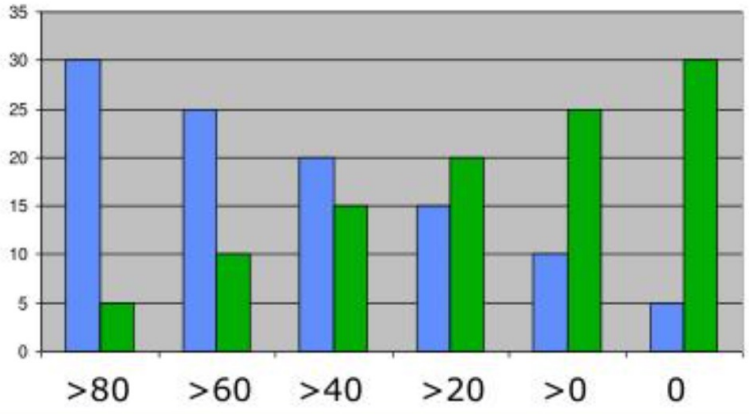
Construction of ROC curve

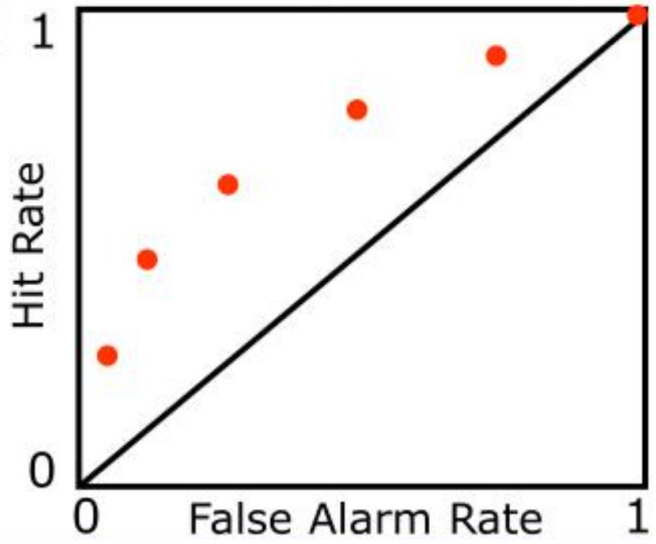
- Determine bins
  - There must be enough occurrences of the event to determine the conditional distribution given occurrences – may be difficult for rare events.
- For each probability threshold, determine HR and FAR
- Plot HR vs FAR to give empirical ROC.



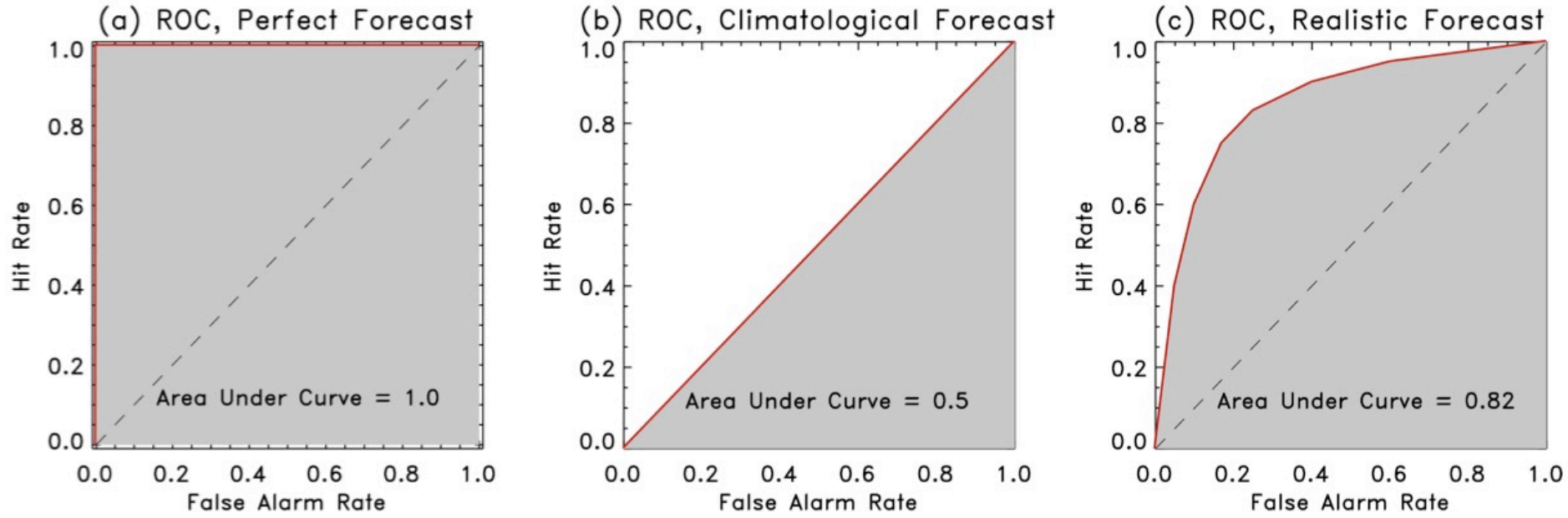
Construction of ROC curve

		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 - Interpretation



- Area under ROC curve (A) used as a single quantitative measure. Area range: 0 to 1. Perfect =1. No Skill = 0.5

- ROC Skill Score (ROCSS)

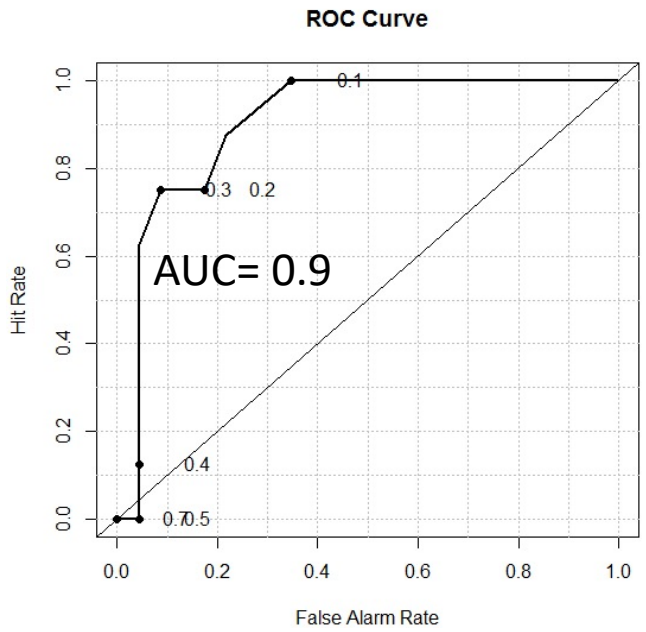
ROCSS = 2A – 1

ROC curve is independent of forecast bias, i.e. represents potential skill

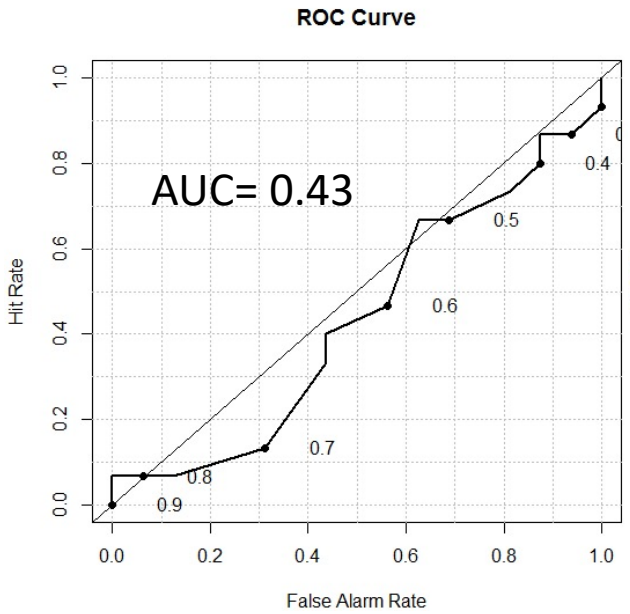
ROC is conditioned on observations (if *y* occurred, what did FC predict?)

Real Example

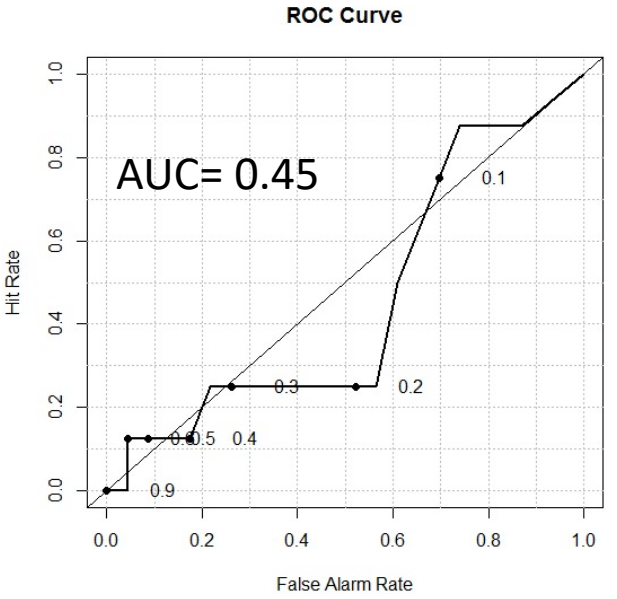
Below Normal



Near Normal



Above Normal



(Mason and Graham, 1999).

Generalized ROC (GROC)

- The GROC, like the ROC, shows the degree of correct probabilistic forecast discrimination, even if the forecasts have biases or calibration problems. However, unlike ROC, GROC is generalized to encompass all forecast categories (below, near, and above normal) collectively, rather than being specific to a single category.
- The GROC is based on a comparison of the forecasts issued for all available pairs of observations of differing category. More specifically, it is the proportion of all such pairs of observations of differing category whose probability forecasts are discriminated in the correct direction.
- The GROC is not plotted as ROC is; however, the score that it produces is fully equivalent to the ROC area, ranging from 0.5 for a set of probability scores without skill to 1 as the maximum possible value.

Comments

- 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,…
- It is important to understand the behaviour of different scores and choose appropriately

References:

- Attributes of Forecast Quality (https://link.springer.com/referenceworkentry/10.1007%2F978-3-642-39925-1_2)
- Descriptions of the IRI Climate Forecast Verification Scores (https://iri.columbia.edu/wp-content/uploads/2013/07/scoredescriptions.pdf)
- Guidance on Verification of Operational Seasonal Climate Forecasts (http://www.seevccc.rs/SEECOF/SEECOF-10/SEECOF-LRF-TRAINING/November%2013th%202013/CCl%20verification%20recommendations.pdf)
- Interpreting the skill score form of forecast performance metrics (https://www.sciencedirect.com/science/article/abs/pii/S0169207019300093)

Thanks!!
Any Question??
npa5302@psu.edu

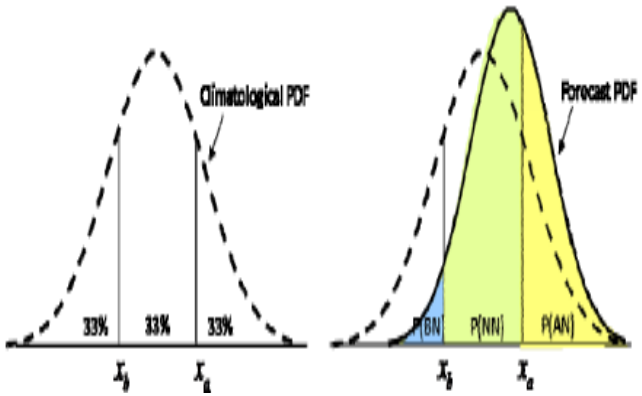
Comments on Rank Histograms

- Not a real verification measure
- Quantification of departure from flatness

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

where RMSD is the root-mean-square difference from flatness, expressed as number of cases, *M* is the total sample size on which the rank histogram is computed, *N* is the number of ensemble members, and S_k is the number of occurrences in the *k*th interval of the histogram.

Calculation of Probabilities of an individual GCM by Parametric



where $X = \beta + \varepsilon$

X is the forecast to be given,

β is the potentially predictable signal

ε is the error part

This yields two more relationship

$$E(X) = E(\beta) \text{ as } E(\varepsilon) = 0$$

$$\sigma_X^2 = \sigma_\beta^2 + \sigma_\varepsilon^2 \text{ as } Cov(\beta, \varepsilon) = 0$$

Forecast Category

$$x_a = -x_b = x_{1/3}\sigma_X = F_N^{-1}(1/3)\sigma_X$$

Probability of Above Normal

$$P_x(AN \mid \beta, \sigma_\varepsilon) = P[(X > x_a) \mid \beta, \sigma_\varepsilon]$$

$$= 1 - F_N\left[\frac{x_a - \beta}{\sigma_\varepsilon}\right]$$

$$= F_N\left[\frac{\beta - x_a}{\sigma_\varepsilon}\right]$$

Probability of Below Normal

$$P_x(BN \mid \beta, \sigma_\varepsilon) = P[(X < x_b) \mid \beta, \sigma_\varepsilon]$$

$$= F_N\left[\frac{x_b - \beta}{\sigma_\varepsilon}\right]$$

Probability of Near Normal

$$P_x(NN \mid \beta, \sigma_\varepsilon) = 1 - P_x(B \mid \beta, \sigma_\varepsilon) - P_x(A \mid \beta, \sigma_\varepsilon)$$

Assumption: Normal distribution

One has to estimate the two unknown parameter β and ε

β = Mean of member

σ_ε = ????

Probabilistic Forecast

Ensemble Spread (ES)

Uncertainty represents by Ensemble spread which is calculated as the variance of ensemble members for a particular year or average of year to year variance of ensemble members.

$$\sigma_t^2 = \frac{1}{M}\sum_{i=1}^{M}\left(f_t^i - \overline{f_t}\right)$$

Error Residual (ER)

Uncertainty represents by Root Mean Square Error (RMSE).

$$MSE = \frac{1}{T}\sum_{t=1}^{T}\left(X_t - \overline{f_t}\right)^2$$

Correlation (CR)

Uncertainty is consider as the function of correlation between observation and signal (β)

$$r = \frac{\mathrm{cov}(Y,\beta)}{\sqrt{\sigma_X{}^2}\,\sigma_\beta^2}$$

$$= \frac{\sigma_\beta}{(\sqrt{(\sigma_\beta{}^2 + \sigma_\varepsilon^2)}}$$

$$\Rightarrow (\sigma_\beta{}^2 + \sigma_\varepsilon^2) = \frac{\sigma_\beta^2}{r^2}$$

$$\Rightarrow \sigma_\varepsilon^2 = \sigma_\beta^2\left(\frac{1}{r^2} - 1\right)$$