



Hydrologic Forecast Verification

Prepared by Satish Regonda,
Julie Demargne and D.-J. Seo

Hydrologic Ensemble Prediction (HEP) Group
Office of Hydrologic Development
NOAA/National Weather Service

RFC Short-Term Ensemble Workshop, November 30, 2006

Sources of this Presentation

- Brown, B.G., (2001), Verification of Probability Forecasts at Points, *WMO QPF Verification Workshop, Prague, Czech Republic*
- Ebert, B., (2003), Verification of Nowcasts, *WWRP Nowcasting Training Workshop, 3-14 November 2003, INMET*
- Ebert, B., (2005), Verification of Ensembles, *TIGGE workshop, 1-3 March 2005, ECMWF*
- Hartmann, H.C., (2006), Hydrologic Forecast Verification
- Hagedorn, R., (2006), EPS Diagnostic Tools, *ECMWF Training Course, Reading*
- Seo, D.-J., (2005), Tutorial Examples of Reliability Diagram and Relative Operating Characteristic from Short-Term Ensemble Prediction, *DSST Ensemble Verification*
- Welles, E., (2004), Hydrology and Verification, *DOH/RDM Conference*

All these presentations are available either online or on OHD server

Types of Forecasts

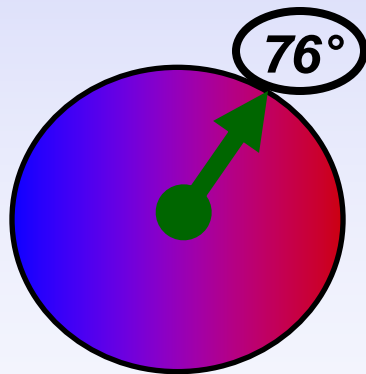
“Today’s high will be **76 degrees**,
and it will be partly cloudy,
with a **30% chance of rain.**”

Deterministic

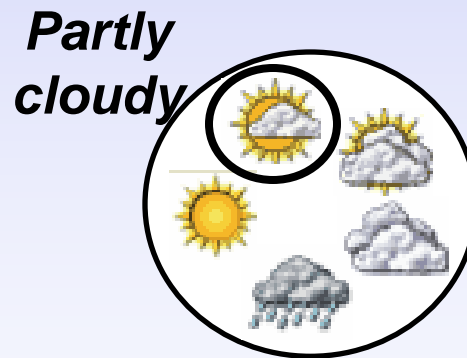
Categorical

Probabilistic

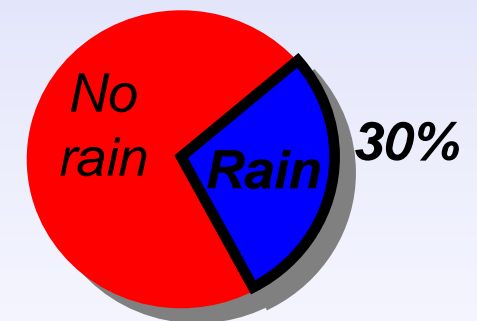
Deterministic



Categorical



Probabilistic



How would you evaluate each of these?

So Many Verification Measures!

- Quantifiable measures

Deterministic

Bias

Correlation

RMSE

Standardized
RMSE

Nash-Sutcliffe

Linear Error in
Probability Space

Skill scores

Categorical

Hit Rate

Surprise rate

Threat Score

Gerrity Score

Success Ratio

Post-agreement

Percent Correct

Pierce Skill Score

Gilbert Skill Score

Heidke Skill Score

Critical Success index

Hannsen and Kuipers Score

Probabilistic

Brier Score

Ranked Probability
Score

Rank Histogram

Distributions-
oriented Measures

Resolution

Reliability

Discrimination

Sharpness

Relative value

So Many Verification Measures!

- Graphical Measures
 - Scatter Plot
 - Rank Histograms
 - Box plot
 - Spread vs. Skill
- Need for easy-to-understand verification measures
- Training to understand how metrics describe different aspects of forecast quality

RFC Verification System: Metrics

CATEGORIES	DETERMINISTIC FORECAST VERIFICATION METRICS	PROBABILISTIC FORECAST VERIFICATION METRICS
1. Categorical <i>(predefined threshold, range of values)</i>	Probability Of Detection (POD), False Alarm Rate (FAR), Lead Time of Detection (LTD), Critical Success Index (CSI), Pierce Skill Score (PSS), Gerrity Score (GS)	Brier Score (BS), Rank Probability Score (RPS)
2. Error <i>(accuracy)</i>	Root Mean Square Error (RMSE), Mean Absolute Error (MAE), Mean Error (ME), Bias (%), Linear Error in Probability Space (LEPS)	Continuous RPS
3. Correlation	Pearson Correlation Coefficient, Ranked correlation coefficient, scatter plots	
4. Distribution Properties	Mean, variance, higher moments for observation and forecasts	Wilcoxon rank sum test, variance of forecasts, variance of observations, ensemble spread, Talagrand Diagram (or Rank Histogram)

RFC Verification System: Metrics

CATEGORIES	DETERMINISTIC FORECAST VERIFICATION METRICS	PROBABILISTIC FORECAST VERIFICATION METRICS
5. Skill Scores <i>(relative accuracy over reference forecast)</i>	Root Mean Squared Error Skill Score (SS-RMSE) (with reference to persistence, climatology, lagged persistence), Wilson Score (WS), Linear Error in Probability Space Skill Score (SS-LEPS)	Rank Probability Skill Score, Brier Skill Score (with reference to persistence, climatology, lagged persistence)
6. Conditional Statistics <i>(based on occurrence of specific events)</i>	Relative Operating Characteristic (ROC), reliability measures, discrimination diagram, other discrimination measures	ROC and ROC Area, other resolution measures, reliability diagram, discrimination diagram, other discrimination measures
7. Confidence <i>(metric uncertainty)</i>	Sample size, Confidence Interval (CI)	Ensemble size, sample size, Confidence Interval (CI)


Why do we need verification measures ?

- Verification statistics help in understanding
 - sources of skill in forecasts
 - sources of uncertainty in forecasts
 - conditions where and when forecasts are skillful or not skillful, and why?
- Also provide information on
 - the accuracy of forecasts
 - the improvement in terms forecast skill and decision making with alternate forecast sources (e.g., climatology, persistence, deterministic forecast)

Thus, helps in finding the limitations (flaws) of the forecast framework and, consequently helps in improving it.

Objective of diagnostic/verification tools

Assess **quality** of forecast system
i.e. determine **skill** and **value** of forecast



A forecast has **skill** if it predicts the observed conditions well according to some objective or subjective criteria.

A forecast has **value** if it helps the user to make better decisions than without knowledge of the forecast.

- Forecasts with poor skill can be valuable (e.g. extreme event forecasted in wrong place)
- Forecasts with high skill can be of little value (e.g. blue sky desert)

What makes an ensemble forecast “good”?

Forecasts should agree with observations, with few large errors

Accuracy

Forecast mean should agree with observed mean

Bias

Linear relationship between forecasts and observations

Association

Forecast should be more accurate than unskilled reference forecasts (e.g., random chance, persistence, or climatology)

Skill

What makes an ensemble forecast “good”?

Binned forecast values should agree with binned observations (agreement between categories)

Reliability

Forecast can discriminate between events & non-events

Resolution

Forecast can predict extreme values

Sharpness

Forecast represents the associated uncertainty

***Spread
(Variability)***

Brier Score (BS)

Brier Score measures mean squared probability error

Consider a specific event by fixing a threshold, then estimate

- p_i , forecast probability, is fraction of members predicting event
- o_i , observed outcome, is “1” if event occurs, otherwise it is “0”

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

BS varies from 0 (perfect deterministic forecasts)
to 1 (perfectly wrong)

Decomposition of BS:

BS = Reliability – Resolution + Uncertainty

Brier Skill Score (BSS)

- **Brier Skill Score (BSS)** measures improvement over reference

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

Positive BSS => better than reference

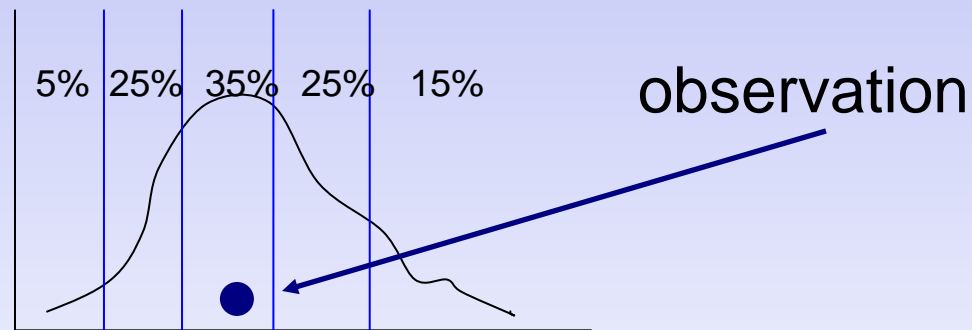
Negative BSS => worse than reference

Analogous to MSE skill score

Ranked Probability Score (RPS)

RPS measures the quadratic distance between forecast and verification probabilities for several categories

Consider multiple events by fixing multiple thresholds



p_i , forecast probability, is fraction of members predicting event

$$p_1=0.05, p_2=0.20, p_3=0.35, p_4=0.25, p_5=0.15$$

o_i , observed outcome, is “1” if event occurs, otherwise it is “0”

$$o_3=1.0, o_1=o_2=o_4=o_5=0.0$$

Ranked Probability Score (RPS)

p_i , forecast probability, is fraction of members predicting event

$$p_1=0.05, p_2=0.20, p_3=0.35, p_4=0.25, p_5=0.15$$

o_i , observed outcome, is “1” if event occurs, otherwise it is “0”

$$o_1=o_2=o_4=o_5=0.0, o_3=1.0$$

$$\text{RPS} = \frac{1}{K-1} \left[\sum_{k=1}^K \left(\sum_{j=1}^k p_j - \sum_{j=1}^k o_j \right)^2 \right]$$

$$\text{RPS} = [(0.05-0.0)^2+(0.25-0.0)^2+(0.60-1.0)^2+(0.85-1.0)^2+(1.0-1.0)^2] / 4$$

$$\text{RPS} = 0.06$$

If $K=2$, then $\text{RPS} = \text{BS}$; Analogous to BS, but multiple categories

Ranked Probability Skill Score (RPSS)

- **RPSS** measures improvement over reference

$$\text{RPSS} = 1 - \frac{\text{RPS}_{\text{for}}}{\text{RPS}_{\text{ref}}}$$

RPS for forecast: RPS_{for}

RPS for reference forecast: RPS_{ref}

RPS for perfect forecast: $\text{RPS}_{\text{per}} = 0$

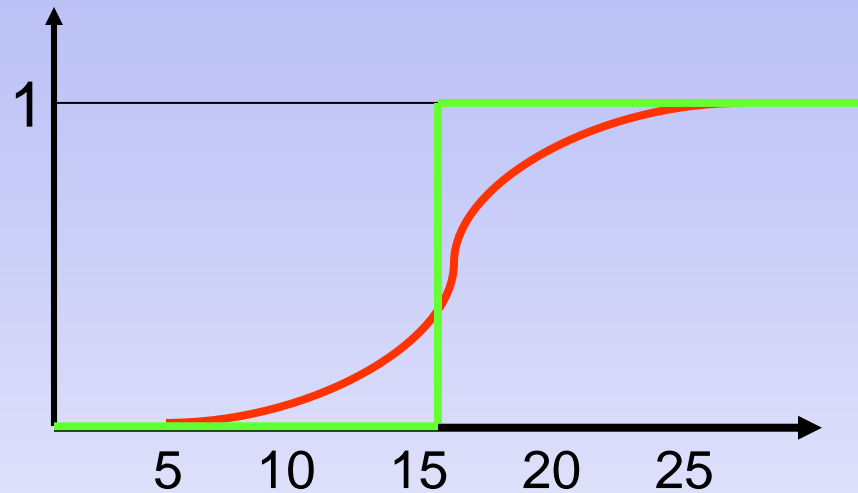
$\text{RPSS} = 0$ \Rightarrow as good as climatology

$\text{RPSS} = 1$ \Rightarrow high skill – much better than climatology

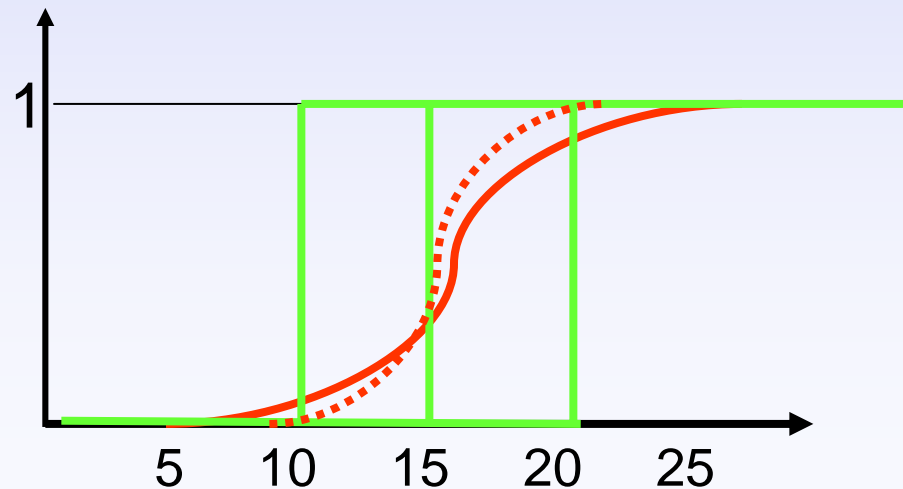
$\text{RPSS} < 0$ \Rightarrow forecast worse than climatology

Brier Score & Ranked Probability Score

- Brier Score used for two category (yes/no) situations (e.g. $T > 15^{\circ}\text{C}$)



- RPS takes into account ordered nature of variable (“extreme errors”)



Source: Hagedorn (2006)

Deterministic Forecast Verification Measures

- **Mean error (bias)** – measures average difference between forecast and observations

$$\text{Mean error} = \frac{1}{N} \sum_{i=1}^N (F_i - O_i)$$

- **Mean Absolute Error (MAE)** – measures average magnitude of forecast error

$$\text{MAE} = \frac{1}{N} \sum_{i=1}^N |F_i - O_i|$$

- **Root Mean Square Error (RMSE)** – measures error magnitude, with large errors having a greater impact than in the MAE

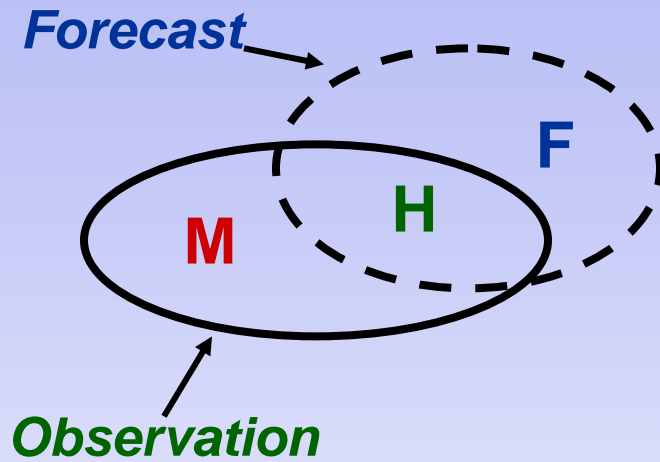
$$\text{RMSE} = \frac{1}{N} \sqrt{\sum_{i=1}^N (F_i - O_i)^2}$$

Deterministic Forecast Verification Measures

- **Pearson Correlation Coefficient** – measures linear correspondence between forecasts and observations

$$r = \frac{\sum (F - \bar{F})(O - \bar{O})}{\sqrt{\sum (F - \bar{F})^2} \sqrt{\sum (O - \bar{O})^2}}$$

Deterministic (Yes/No) Forecast Verification Measures



H = Hits

M = Misses

F = False Alarms

Probability of Detection (POD) – measures fraction of events that were correctly forecast to occur

$$\text{POD} = H / (H+M)$$

False Alarm Ratio (FAR) – measures fraction of "yes" forecasts that were incorrect

$$\text{FAR} = F / (F+H)$$

BIAS score – measures ratio of forecast frequency to observed frequency

$$\text{BIAS} = (H+F) / (H+M)$$

Verification of two category (yes/no) situation

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

		Event observed		total
		Yes	No	
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: $B = (a+b) / (a+c)$
- Hit Rate: $H = a / (a+c)$
- False Alarm Rate: $F = b / (b+d)$
- False Alarm Ratio: $FAR = b / (a+b)$

Example of Finley Tornado Forecasts (1884)

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

		Event observed		total
		Yes	No	
Event forecasted	Yes	28	72	100
	No	23	2680	2703
total		51	2752	2803

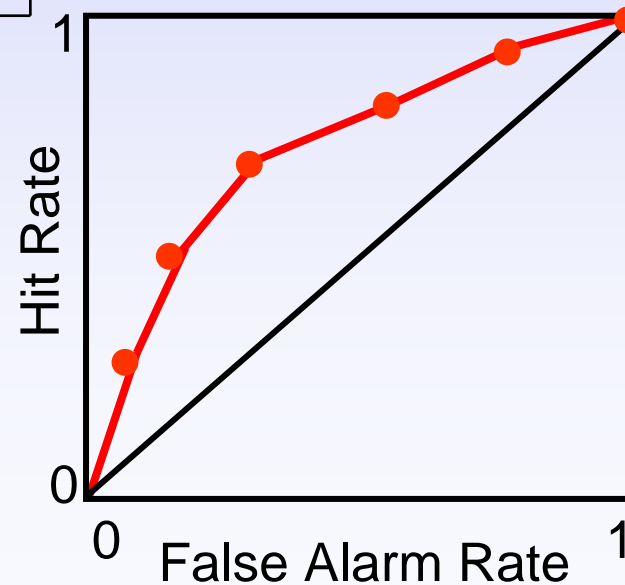
- Event Probability: $s = (a+c) / n = 0.018$
- Probability of a Forecast of occurrence: $r = (a+b) / n = 0.036$
- Frequency Bias: $B = (a+b) / (a+c) = 1.961$
- Hit Rate: $H = a / (a+c) = 0.549$
- False Alarm Rate: $F = b / (b+d) = 0.026$
- False Alarm Ratio: $FAR = b / (a+b) = 0.720$

Extension of 2 x 2 Contingency Table for Probabilistic Forecast

		Event observed		<i>threshold</i>	<i>H</i>	<i>F</i>
		<i>Yes</i>	<i>No</i>			
Event forecasted	>80% - 100%	30	5	>80%	30/105	5/105
	>60% - 80%	25	10	>60%	55/105	15/105
	>40% - 60%	20	15	>40%	75/105	30/105
	>20% - 40%	15	20	>20%	90/105	50/105
	>0% - 20%	10	25	>0%	100/105	75/105
	0%	5	30		105/105	105/105
	total	105	105			

Extension of 2 x 2 Contingency Table for Probabilistic Forecast

		Event observed		<i>threshold</i>	<i>H</i>	<i>F</i>
		Yes	No			
Event forecasted	>80% - 100%	30	5	>80%	0.29	0.05
	>60% - 80%	25	10	>60%	0.52	0.14
	>40% - 60%	20	15	>40%	0.71	0.29
	>20% - 40%	15	20	>20%	0.86	0.48
	>0% - 20%	10	25	>0%	0.95	0.71
	0%	5	30		1.00	1.00
total		105	105			

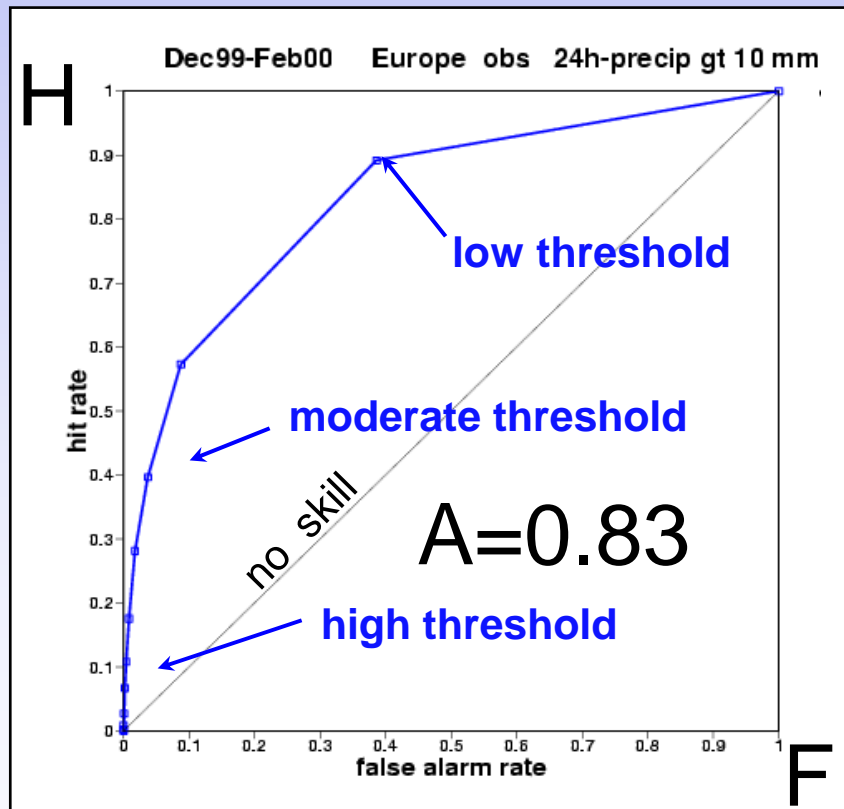


Source: Hagedorn (2006)

Relative Operating Characteristic (ROC)

ROC measures the ability of forecast to discriminate between events and no-events

ROC curve: plot of H against F for range of probability thresholds



ROC area: area under the ROC curve; measures skill

$A=0.5 \Rightarrow$ no skill

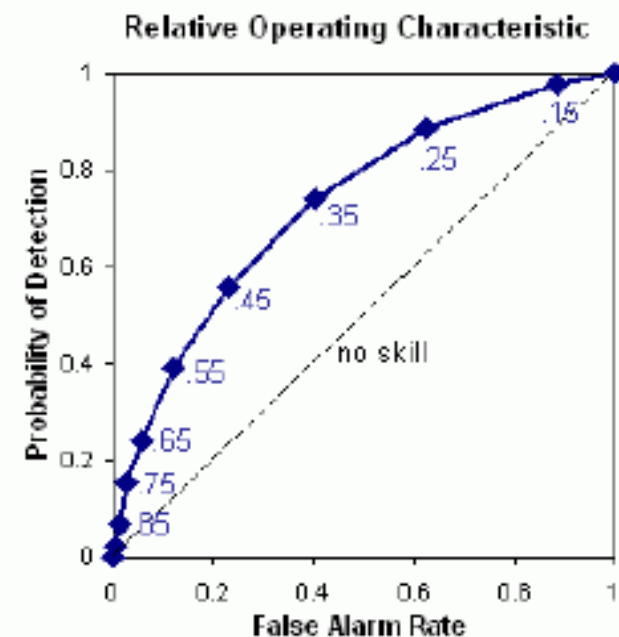
$A=1 \Rightarrow$ perfect deterministic forecast

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.

- 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 = $2(\text{ROC area} - 0.5)$
 - Not sensitive to bias.
 - The ROC is conditioned on the observations (i.e., given that Y occurred, what was the corresponding forecast?)
 - Reliability and ROC diagrams are good companions



Comparison of Approaches

- Brier score
 - Based on squared error
 - Strictly proper scoring rule
 - Calibration is an important factor; lack of calibration impacts scores
 - Decompositions provide insight into several performance attributes
 - Dependent on frequency of occurrence of the event
- ROC
 - Considers forecasts' ability to discriminate between Yes and No events
 - Calibration is not a factor
 - Less dependent on frequency of occurrence of event
 - Provides verification information for individual decision thresholds

Reliability Diagrams

“When you say 80% chance of high flows, how often do high flows happen?”

$P(O/F)$

Graphical Representation of Measures

Discrimination

- Do the forecasts discriminate between types of future events?
- If a flood happened was there a forecast?
- Sort based upon the observed values

=> Discrimination diagram

$p(f/x=0)$ and $p(f/x=1)$

Reliability

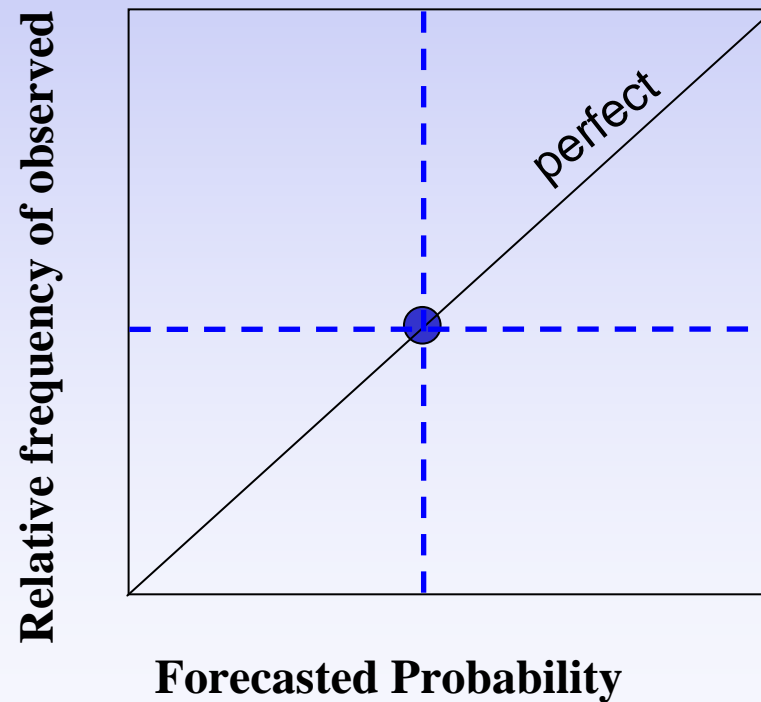
- When we forecast an event, are the forecasts reliable?
- If we forecast some thing, does it happen?
- Sort based upon the forecast values

=> Reliability diagram

$p(x=1|f_i)$ vs. f_i

Forecast Reliability

*If the forecast says there's a 50% chance of **wet**,
wet should happen 50% of the time*



Forecast Reliability

Reliability Diagrams

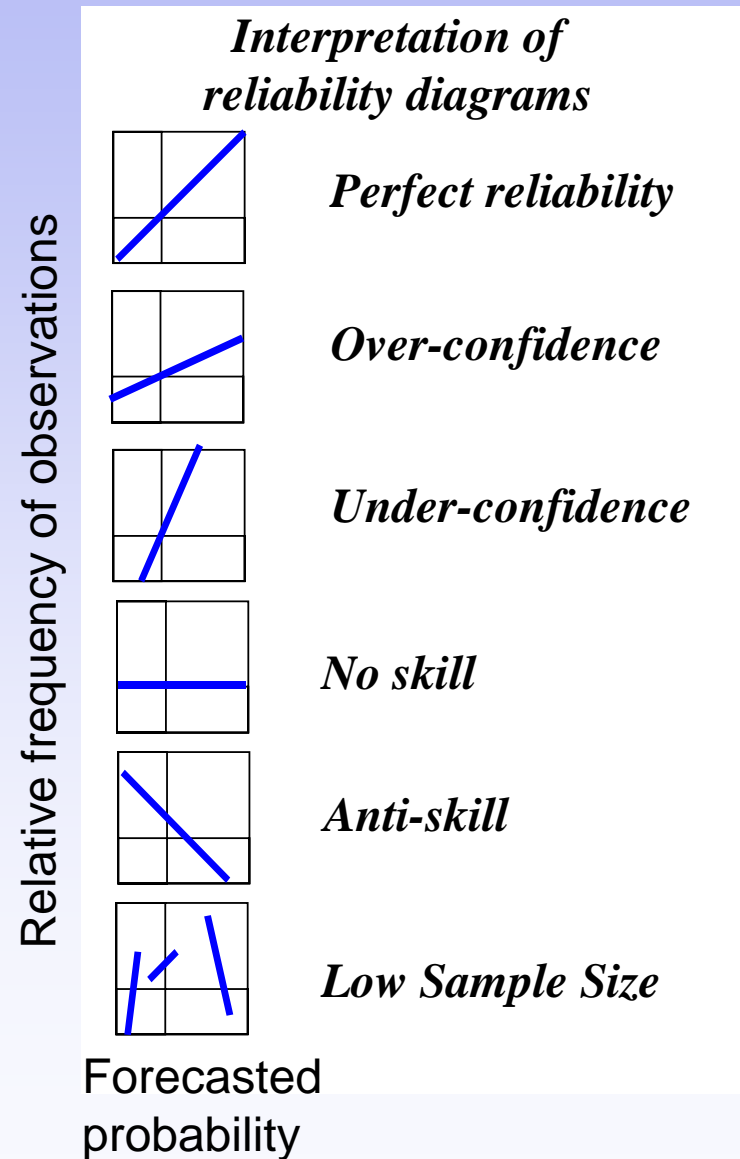
identifies conditional bias

Reliability

$P[O|F]$

Does the frequency of occurrence match your probability statement?

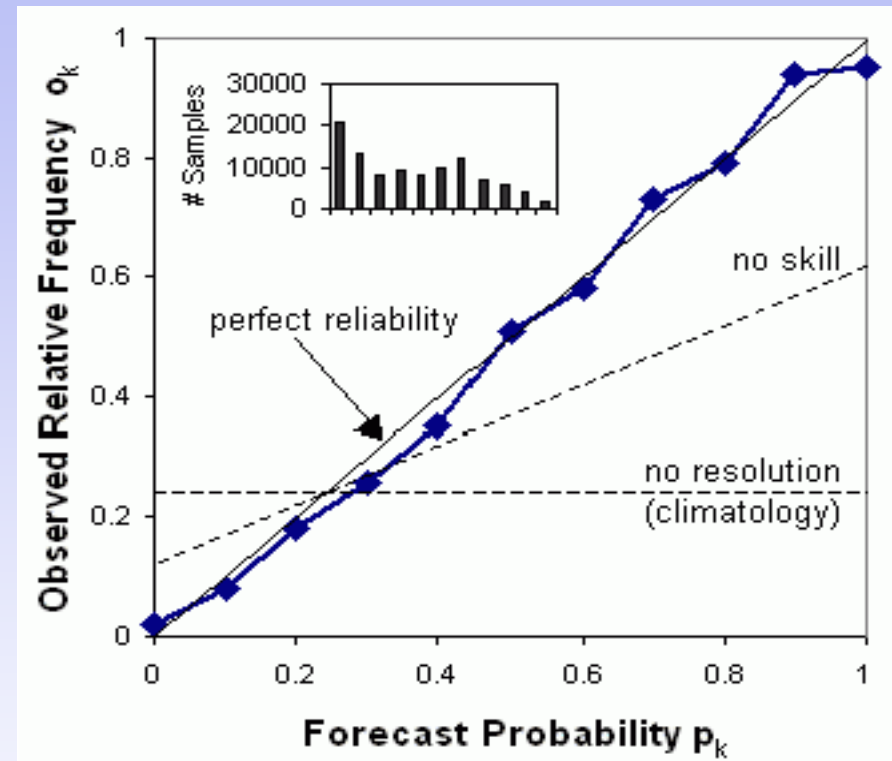
Source: Hartmann (2006)



Reliability (Attribute) Diagram

Attributes diagram: Reliability, Resolution, Skill/No-skill

- Good reliability – close to diagonal
- Good resolution – wide range of frequency of observations corresponding to forecast probabilities
- Sharpness diagram ($p(f)$) – histogram of forecasts in each probability bin shows the sharpness of the forecast.



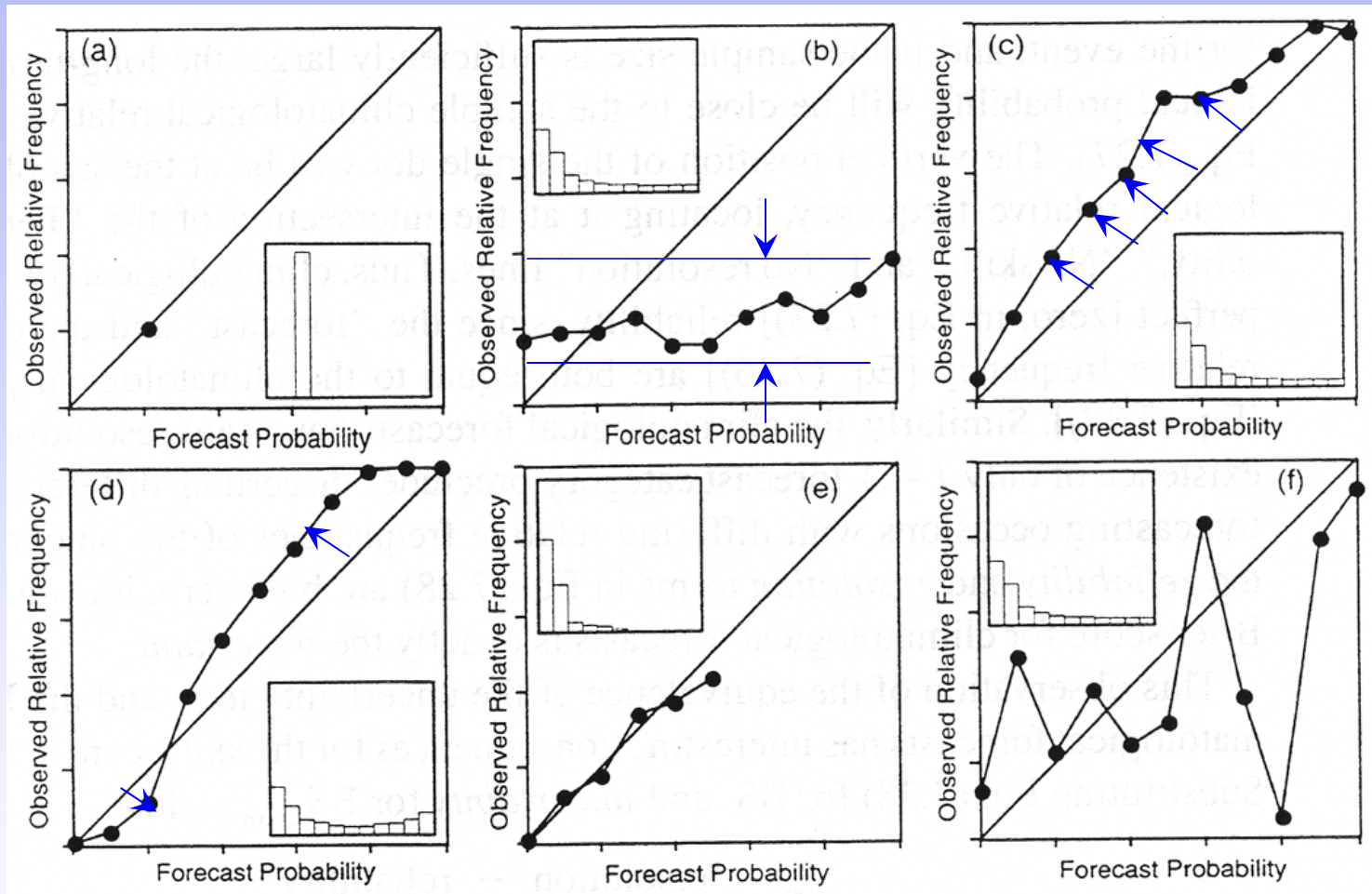
The reliability diagram is conditioned on the forecasts (i.e., given that X was predicted, what was the outcome?), and can be expected to give information on the real meaning of the forecast.

Reliability and Sharpness

Climatology

Minimal RESolution

Underforecasting



**Good RES, at
expense of REL**

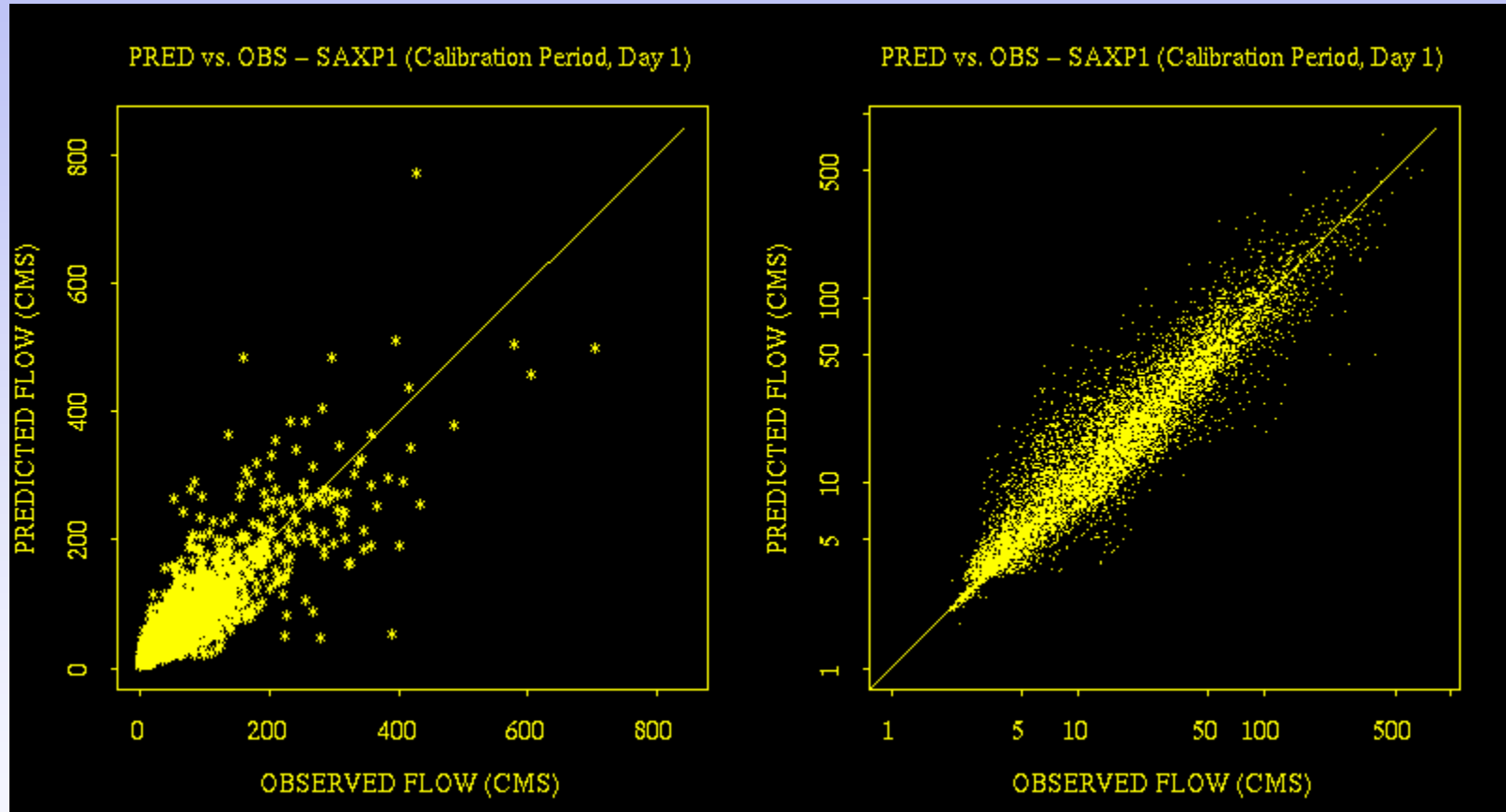
**Reliable forecasts
of rare event**

Small sample size

Examples: 1) Good Reliability

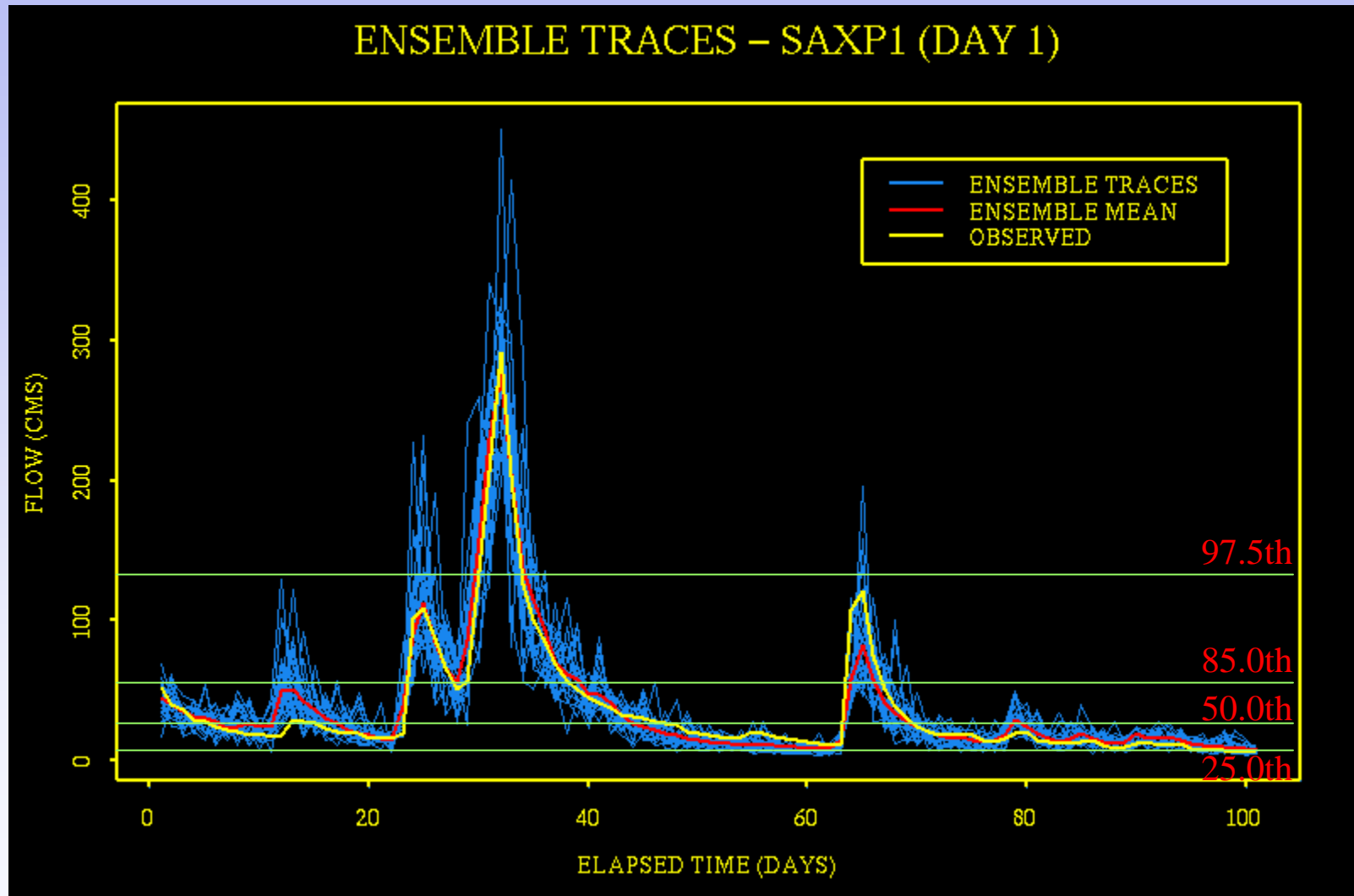
Simulated flows

Postprocessed flows



Source: Seo (2005)

Examples: 1) Good Reliability

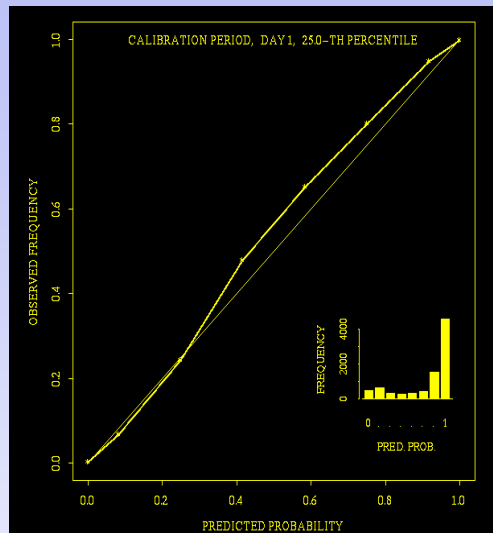


Source: Seo (2005)

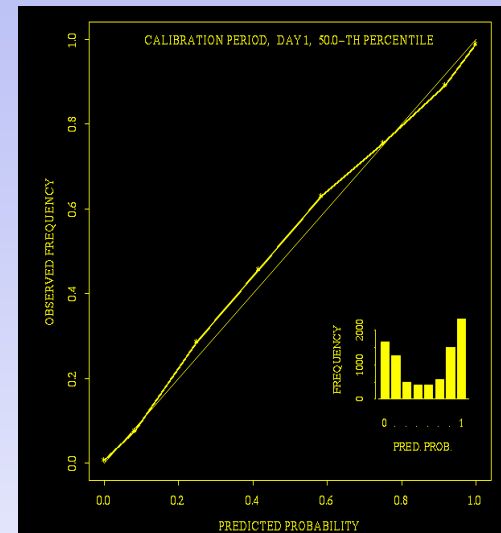
Examples: 1) Good Reliability

Reliability Diagram (agreement between forecast probability and mean observed frequency)

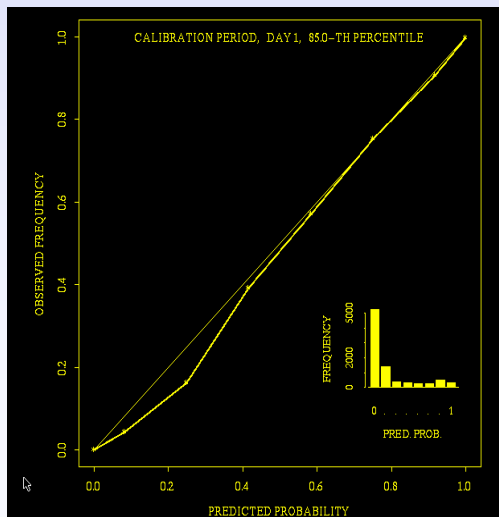
25th percentile



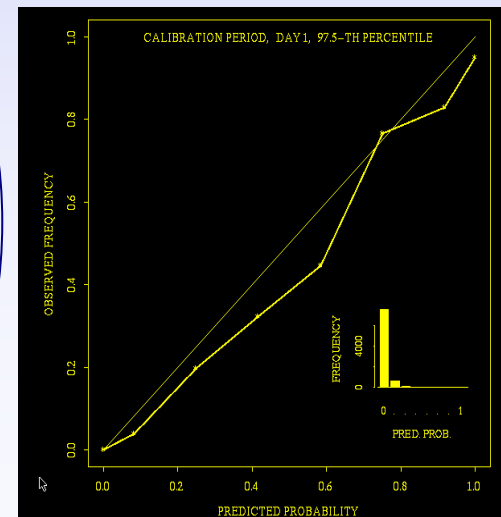
50th percentile



85th percentile



97.5th percentile

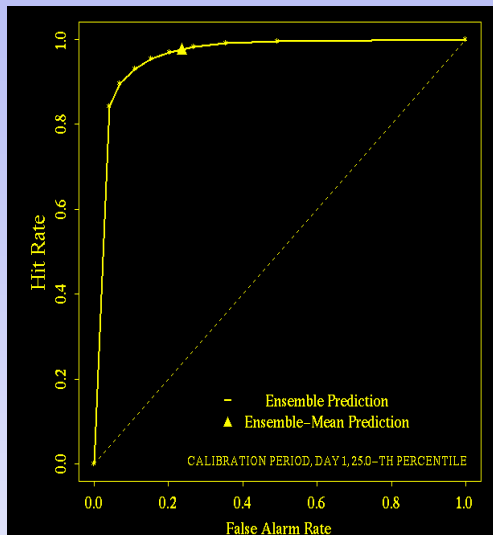


Source: Seo (2005)

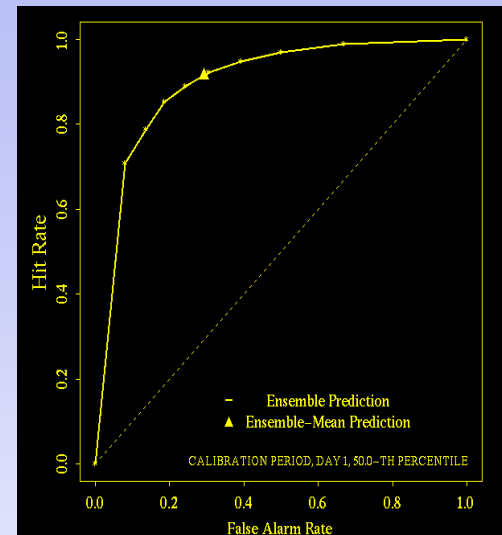
Examples: 1) Good Reliability

ROC (ability of forecast to discriminate between events & non-events)

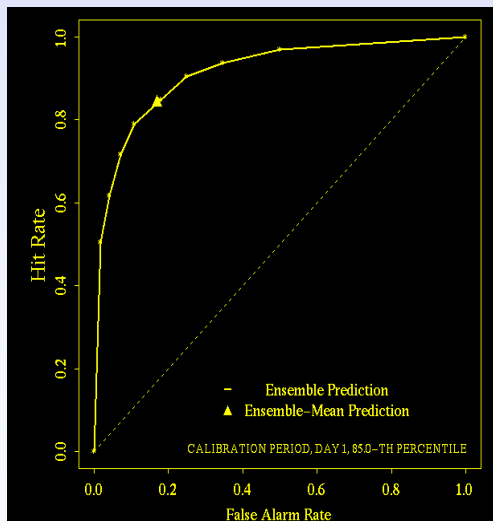
25th percentile



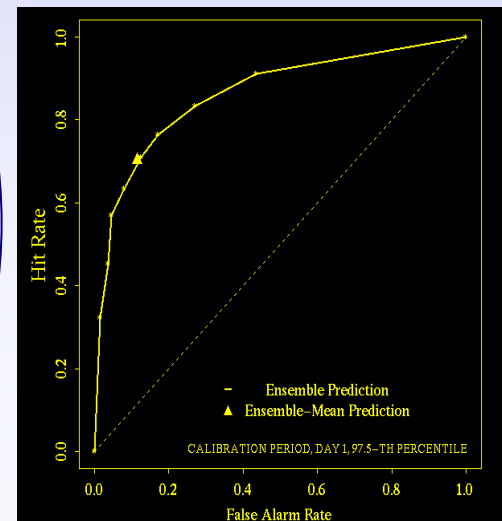
50th percentile



85th percentile



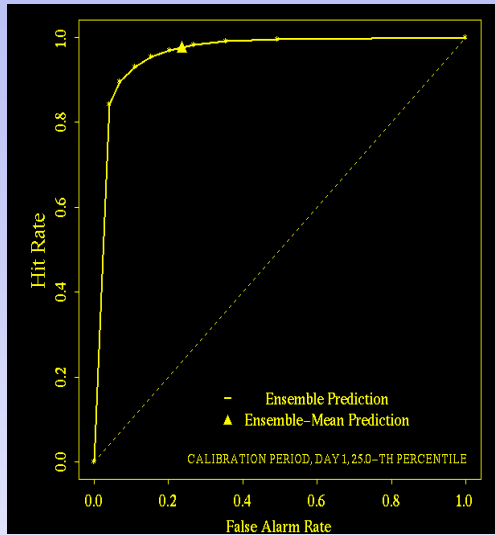
97.5th percentile



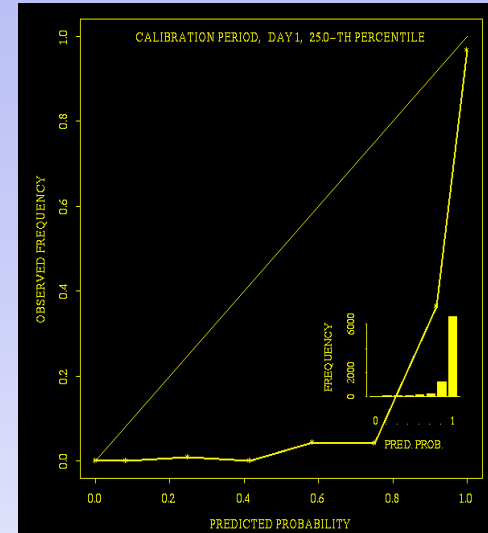
Source: Seo (2005)

Examples: ROC

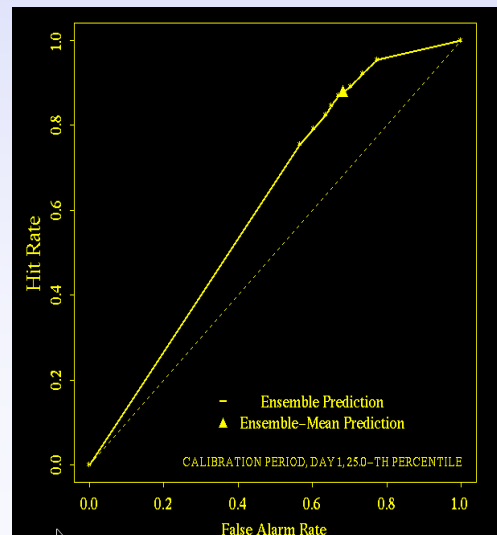
ROC (ability of forecast to discriminate between events & non-events)



Good Forecast



Over estimated

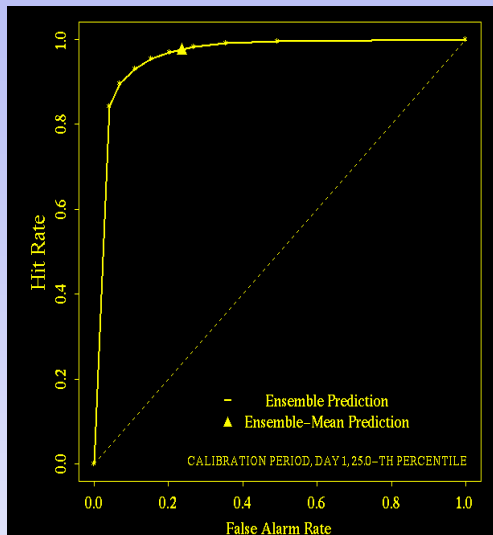


Random (no skill)

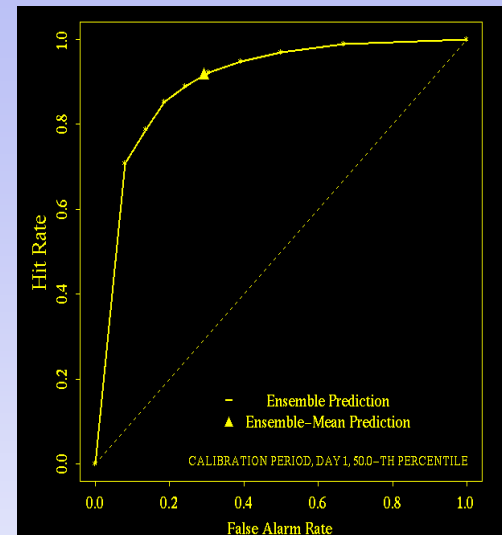
Examples: 1) Good Reliability

ROC (ability of forecast to discriminate between events & non-events)

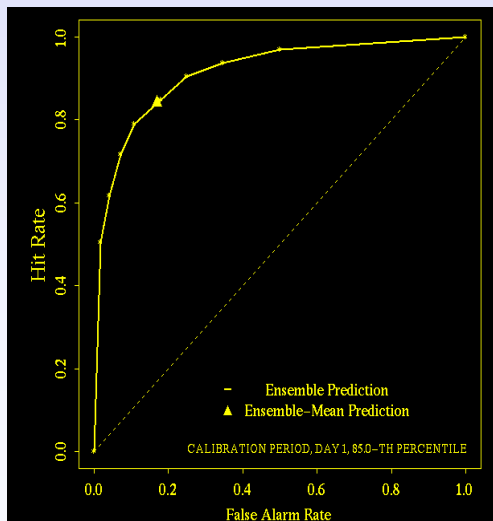
25th percentile



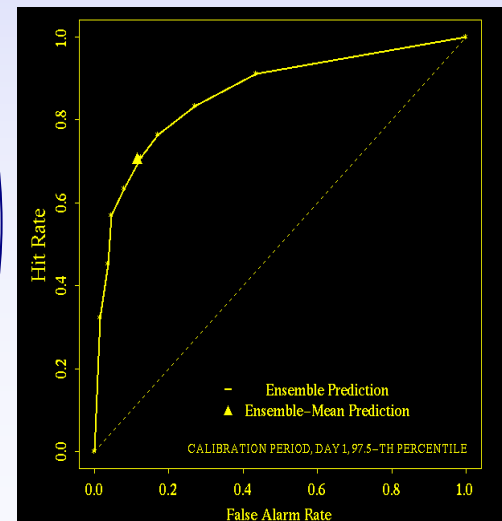
50th percentile



85th percentile



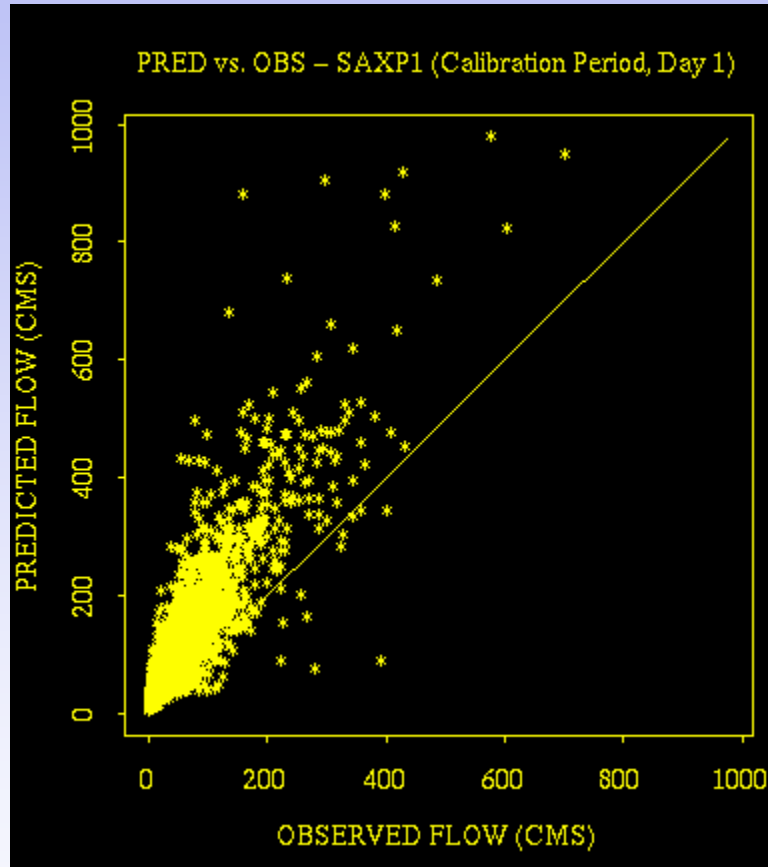
97.5th percentile



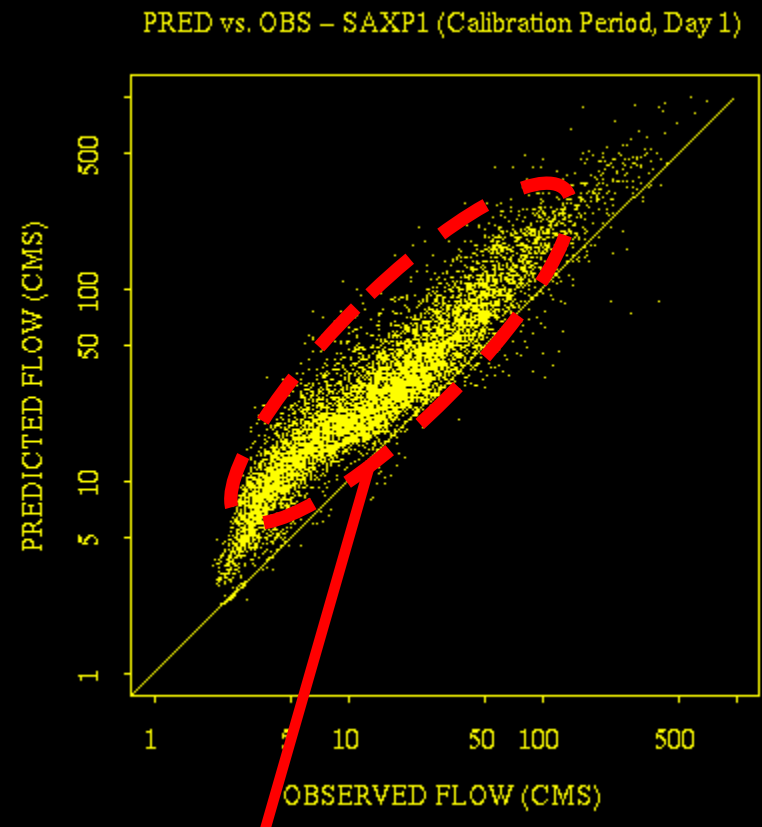
Source: Seo (2005)

Examples: 2) Positive Bias / Overestimated

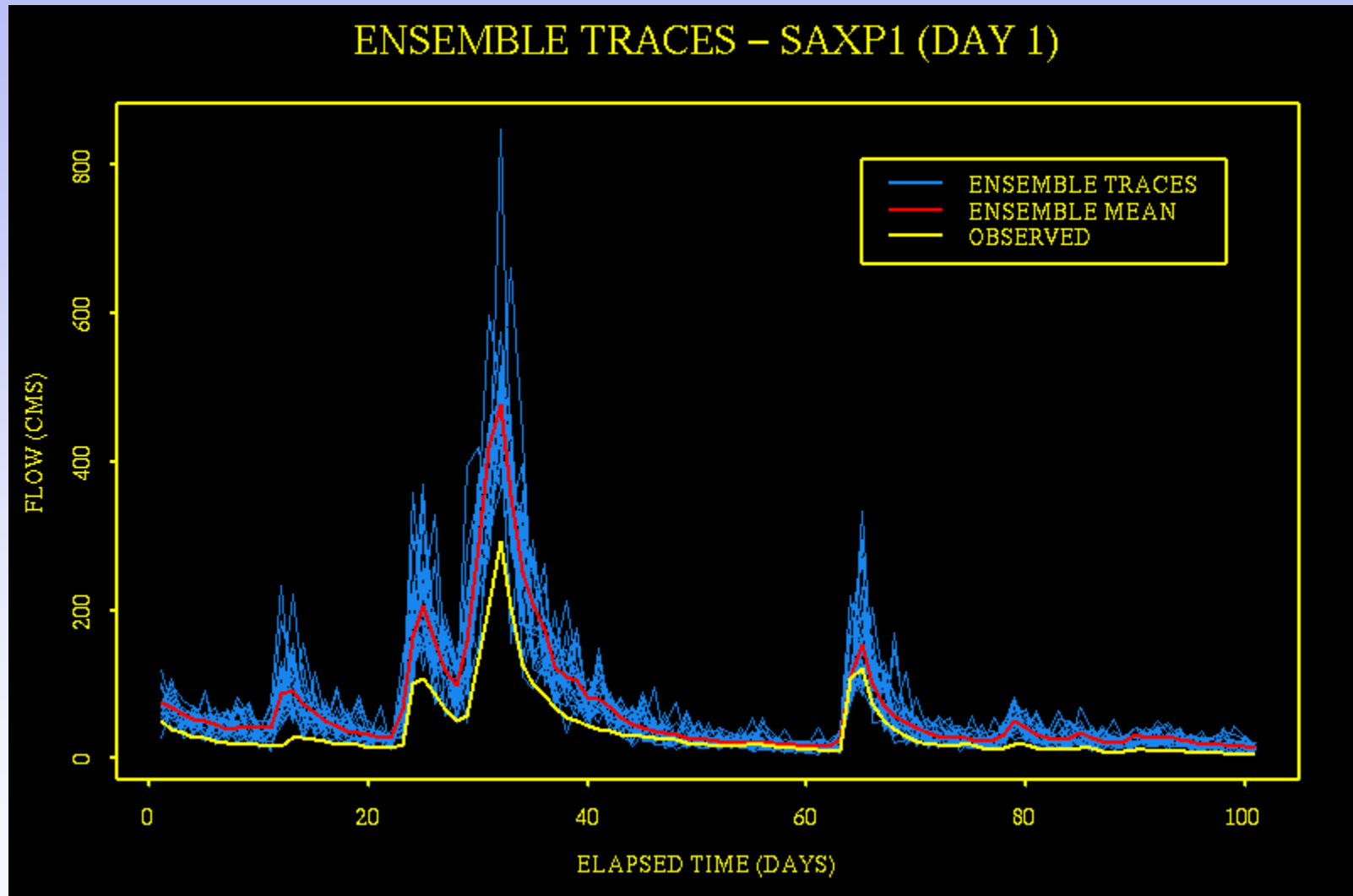
Simulated flows



Postprocessed flows



Examples: 2) Positive Bias / Overestimated

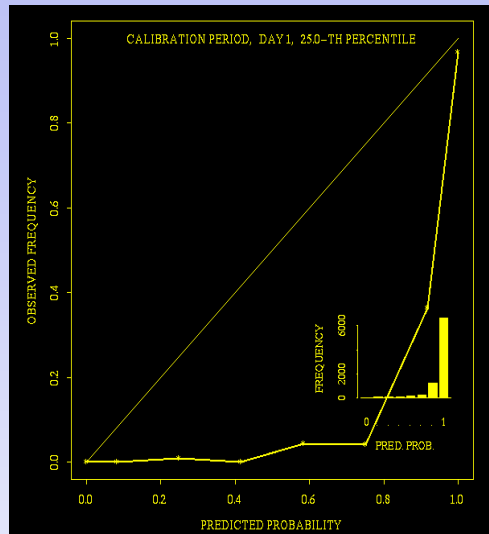


Source: Seo (2005)

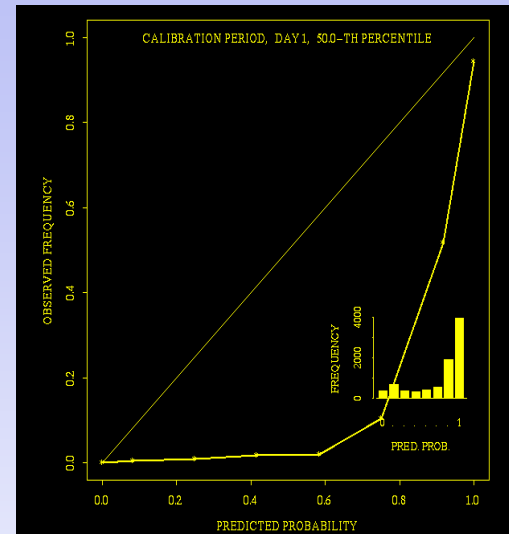
Examples: 2) Positive Bias / Overestimated

Reliability Diagram (agreement between forecast probability and mean observed frequency)

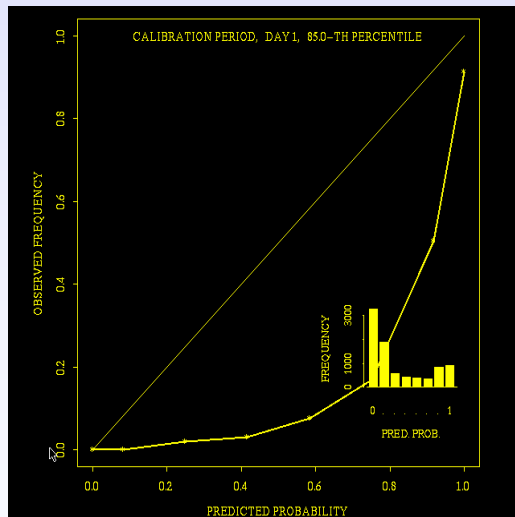
25th percentile



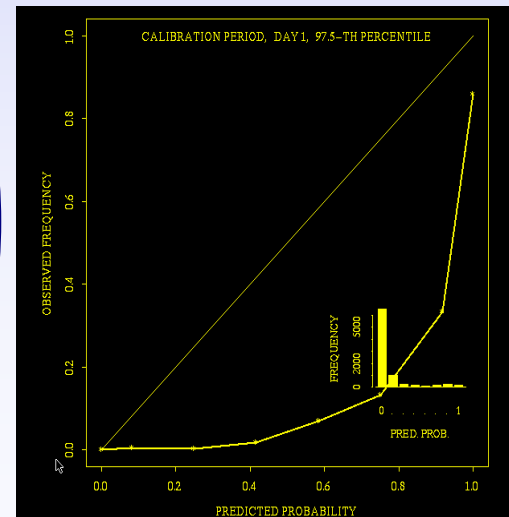
50th percentile



85th percentile



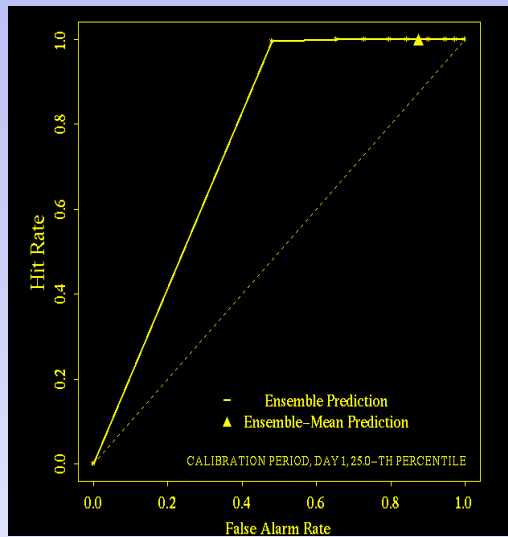
97.5th percentile



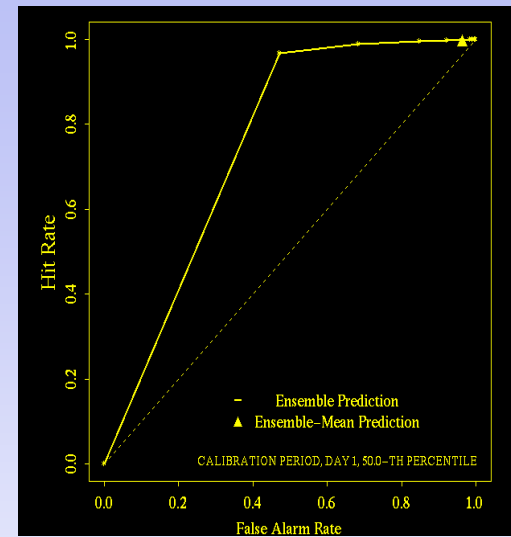
Examples: 2) Positive Bias / Overestimated

ROC (ability of forecast to discriminate between events & non-events)

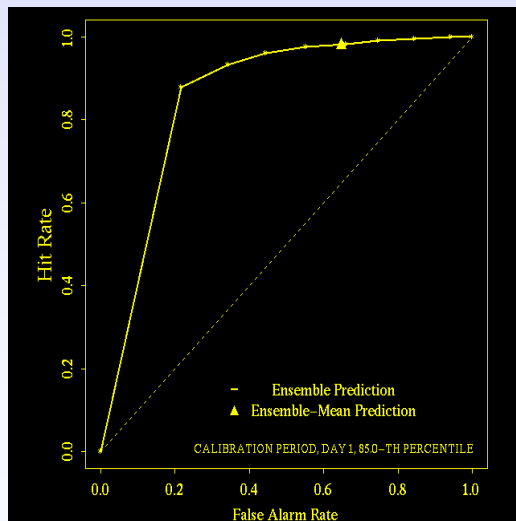
25th percentile



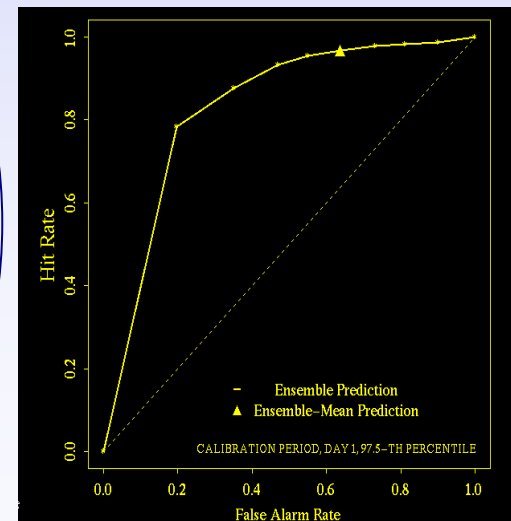
50th percentile



85th percentile



97.5th percentile

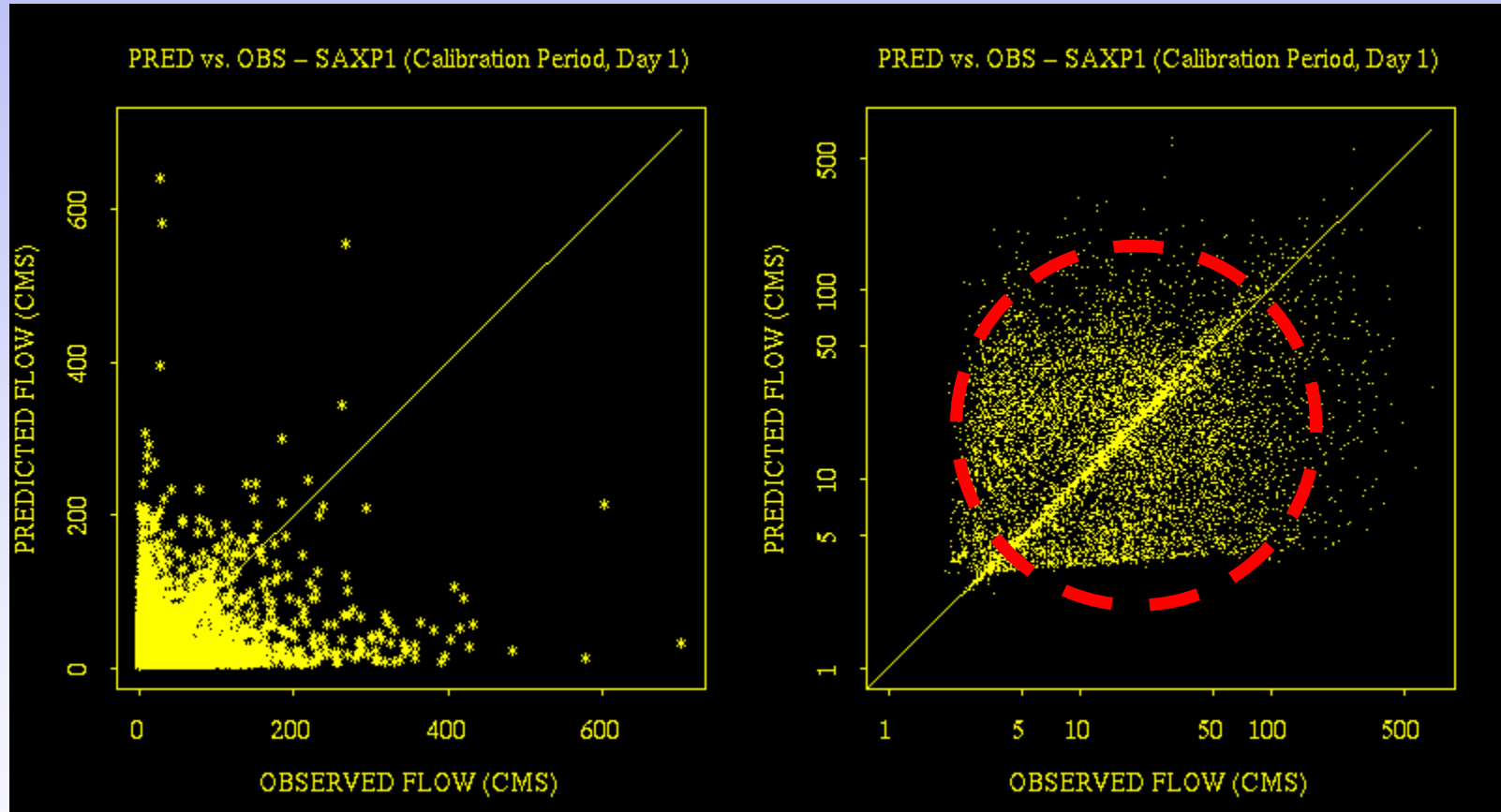


Source: Seo (2005)

Examples: 3) No Skill / Random

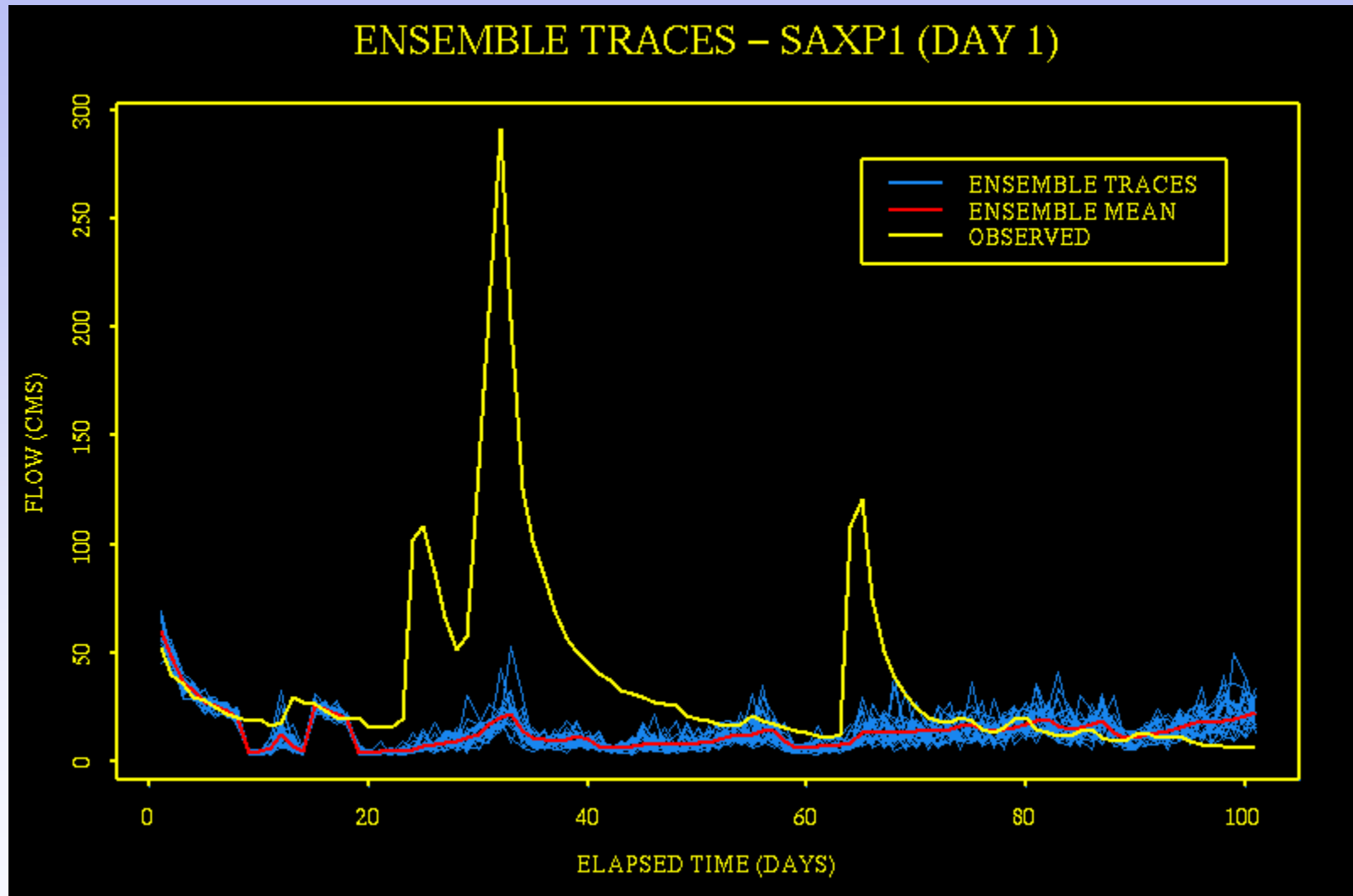
Simulated flows

Postprocessed flows



Source: Seo (2005)

Examples: 3) No Skill / Random

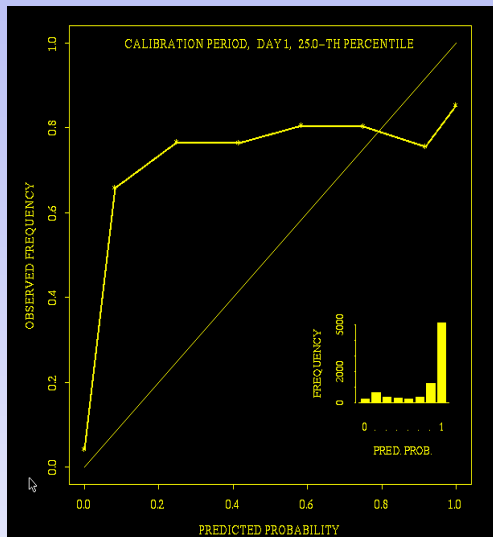


Source: Seo (2005)

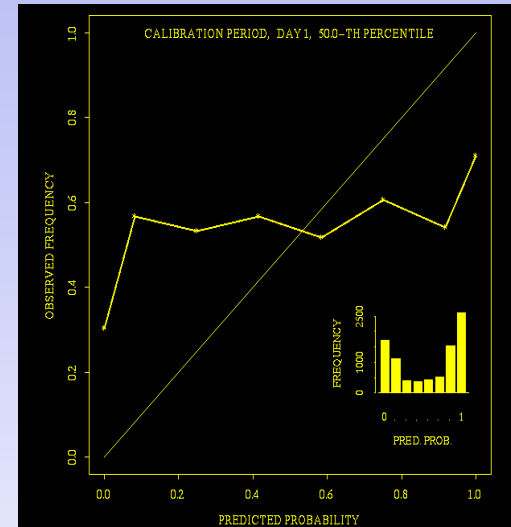
Examples: 3) No Skill / Random

Reliability Diagram (agreement between forecast probability and mean observed frequency)

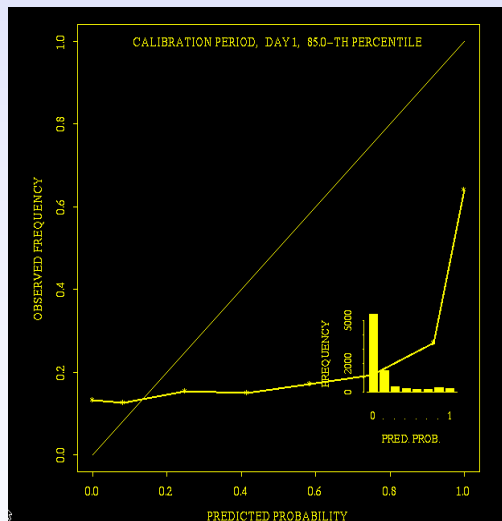
25th percentile



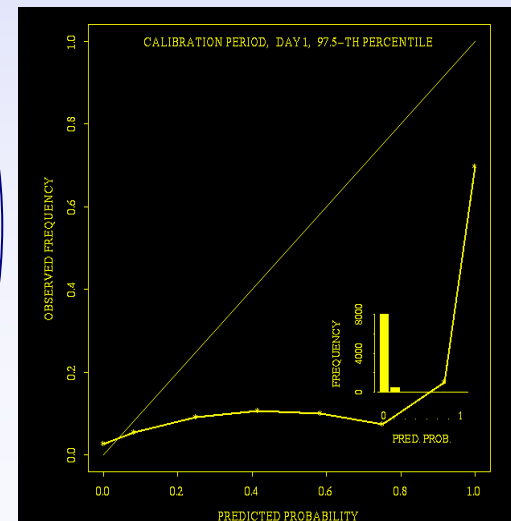
50th percentile



85th percentile



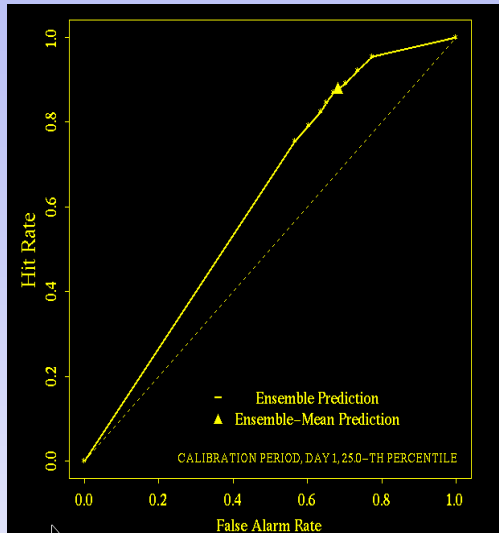
97.5th percentile



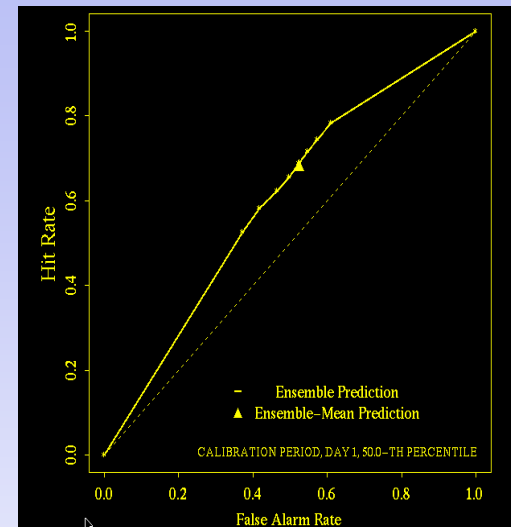
Examples: 3) No Skill / Random

ROC (ability of forecast to discriminate between events & non-events)

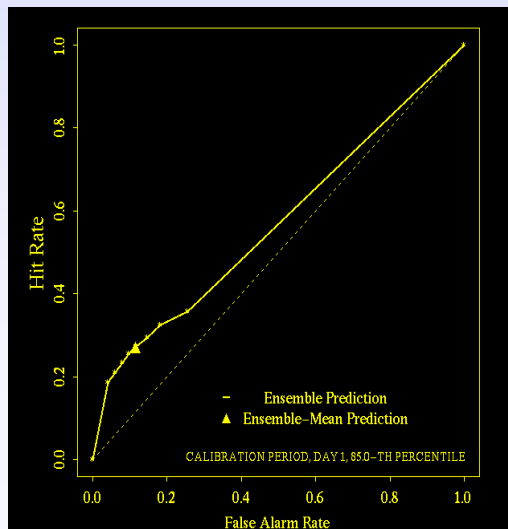
25th percentile



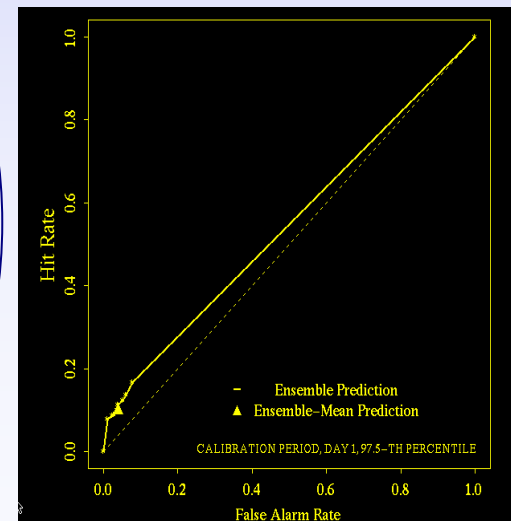
50th percentile



85th percentile



97.5th percentile



Source: Seo (2005)

Discrimination Diagrams

**“When dry happened,
what were the forecasts up to?”**

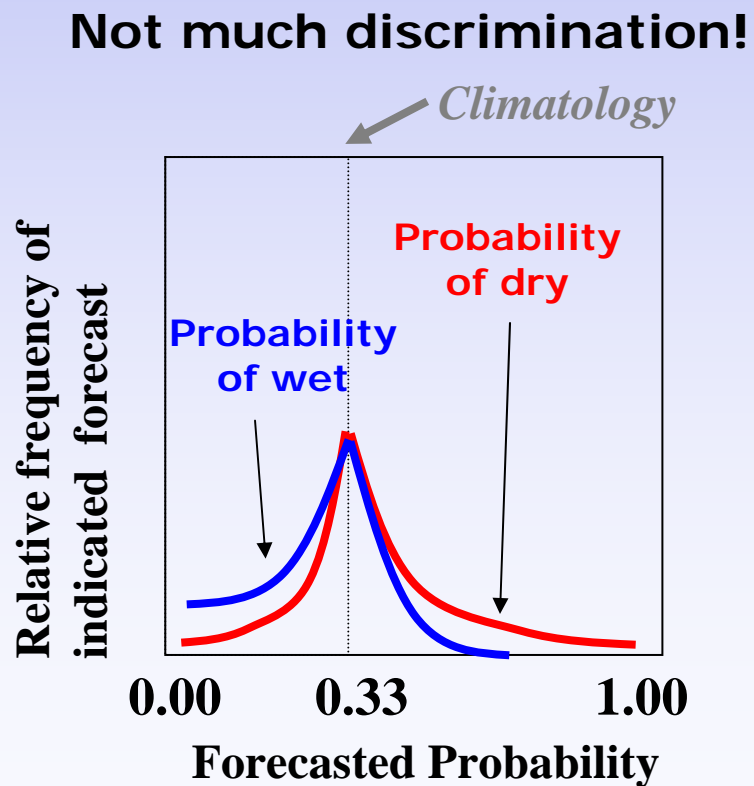
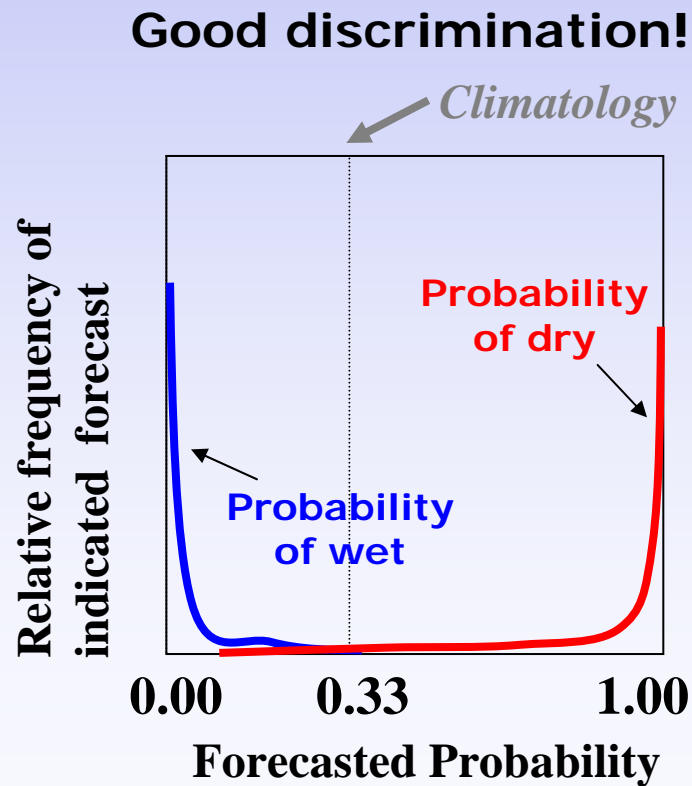
$P(F|O)$

Discrimination Diagrams

“When **dry** happened, what were the forecasts up to?”

Discrimination: $P[F|O]$

Can the forecasts distinguish among different events?



Source: Hartmann (2006)

Thank you