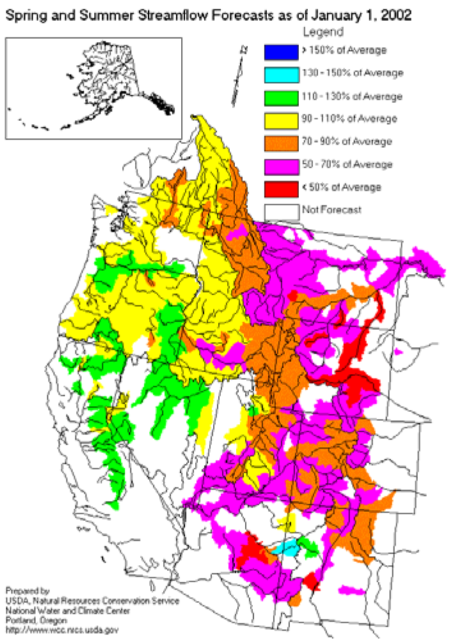
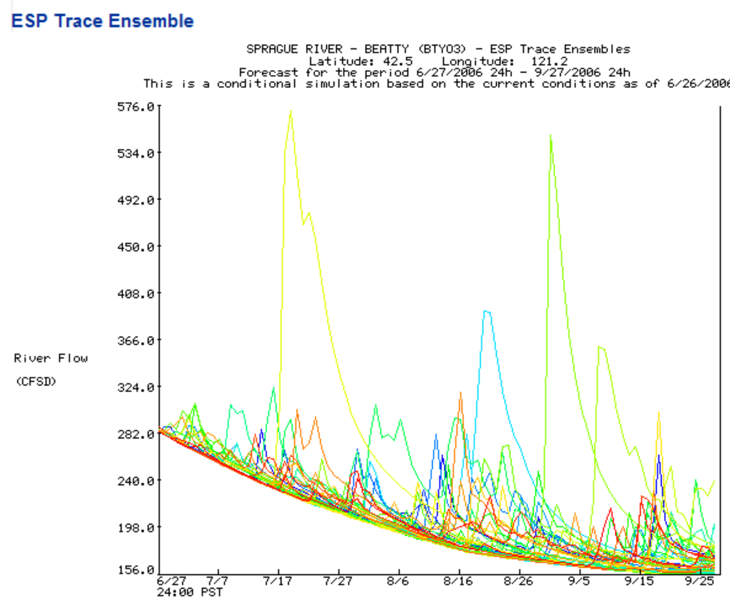
Supported by:

Topics in Ensemble Verification





Holly C. Hartmann

The University of Arizona

hollyoregon@juno.com

NOAA GAPP

NOAA CLIMAS

HyDIS: NASA/Raytheon Synergy

NSF SAHRA

NWS CSD

Stakeholder Use of HydroClimate Info & Forecasts

Common across all groups

Uninformed, mistaken about forecast interpretation

Use of forecasts limited by lack of demonstrated forecast skill

Have difficulty specifying required accuracy

Common across many, but not all, stakeholders

Have difficulty distinguishing between "good" & "bad" products

Have difficulty placing forecasts in historical context

Unique among stakeholders

Relevant forecast variables, regions (location & scale), seasons, lead times, performance characteristics

Technical sophistication: base probabilities, distributions, math

Role of of forecasts in decision making

RFC Verification Priorities: Metrics

CATEGORIES	DETERMINISTIC FORECAST VERIFICATION METRICS	PROBABILISTIC FORECAST VERIFICATION METRICS
1. Categorical ***(predefined threshold, range of values)***	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 ***(accuracy)***	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)**

Source: Verification Group, courtesy J. Demargne

RFC Verification Priorities: Metrics

CATEGORIES	DETERMINISTIC FORECAST VERIFICATION METRICS	PROBABILISTIC FORECAST VERIFICATION METRICS
5. Skill Scores ***(relative accuracy over reference forecast)***	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 ***(based on occurrence of specific events)***	Relative Operating Characteristic (ROC), reliability measures, discrimination diagram**,** **other discrimination measures**	ROC and ROC Area**,** reliability diagram, discrimination diagram**,** **other discrimination measures**
7. Confidence ***(metric uncertainty)***	Sample size**,** **Confidence Interval (CI)**	**Ensemble size,** sample size, **Confidence Interval (CI)**

Source: Verification Group, courtesy J. Demargne

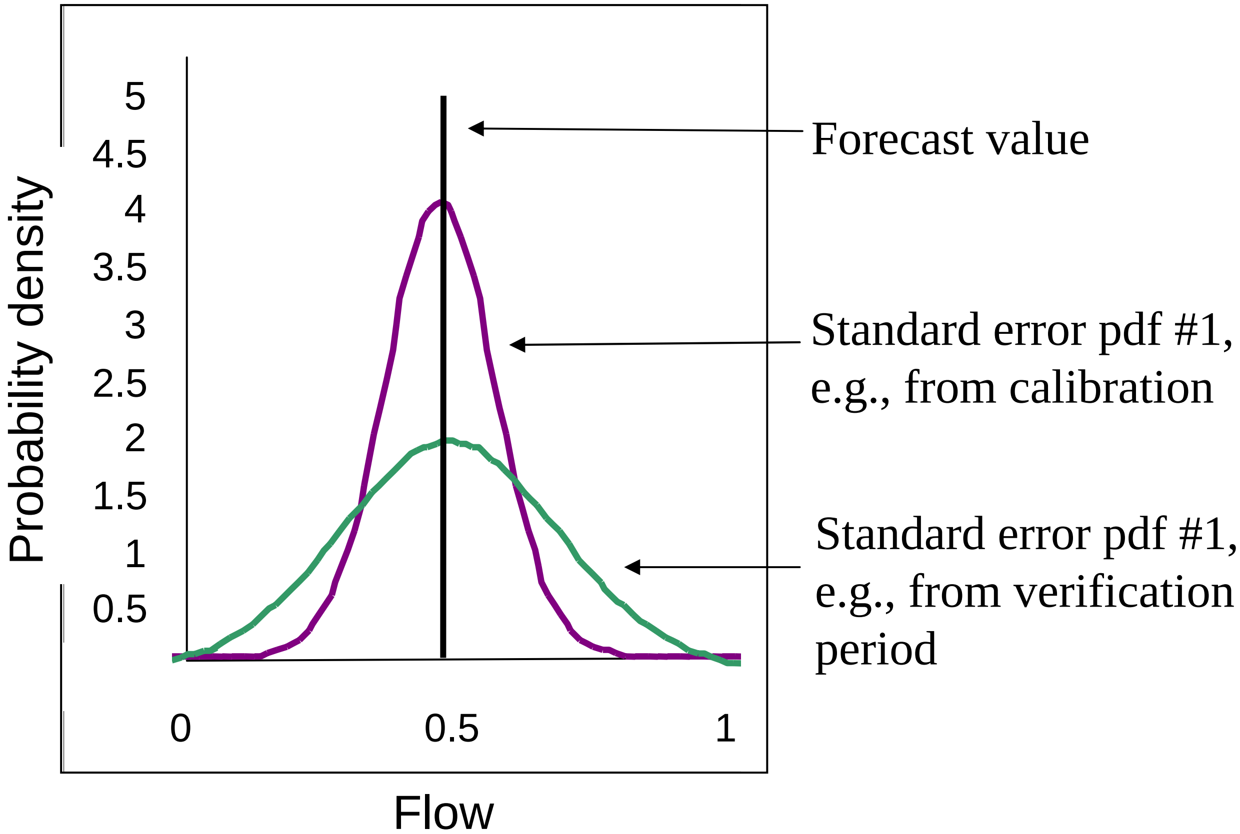
COMET Training: First Module on Verification

INTRODUCTION TO
VERIFICATION OF HYDROLOGIC FORECASTS

Topic	Measures: Deterministic	Measures: Probabilistic	Section
Distribution Properties	Mean Variance Standard Deviation PDF, CDF IQR	PDF, CDF IQR Rank Histogram	2
Forecast Confidence	Sample Size, Confidence Interval	Sample Size, Confidence Interval	3
Correlation	Scatter Plots Correlation Coefficient		4
Categorical Forecast Statistics	Probability of Detection (POD) False Alarm Ratio (FAR) Probability of False Detection (POFD) Bias Critical Success Index (CSI)	Brier Score (BS) Ranked Probability Score (RPS)	5
Accuracy (Error Statistics)	Mean Absolute Error (MAE) Root Mean Square Error (RMSE) Mean Error (ME) Volumetric Bias	Continuous RPS (CRPS)	6
Forecast Skill	Root Mean Square Error Skill Score (RMSE-SS)	Brier Skill Score (BSS) Ranked Probability Skill Score (RPSS)	7
Conditional Verification	Reliability Measures Relative Operating Characteristic (ROC)	Reliability Diagram Attributes Diagram Discrimination Diagram Relative Operating Characteristic (ROC)	8

Deterministic Forecast PDF

Forecast value with standard error, e.g., from calibration, from long-term verification activities



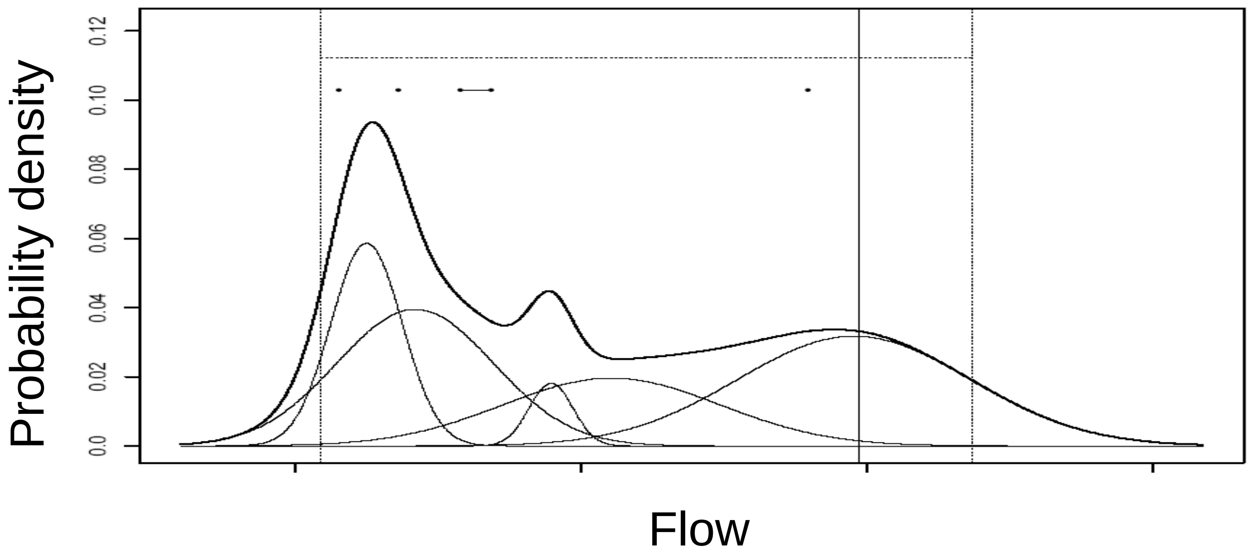
What would the pdf look like for a perfect forecast system?

How would you interpret this forecast for a user?

How do these conceptual examples differ from real-world pdfs?

Ensemble Forecast PDF

Hydrology forecasts: Based on mixed distribution of meteorological probabilities: no precipitation, extreme precipitation, snow, rain-on-snow, etc.



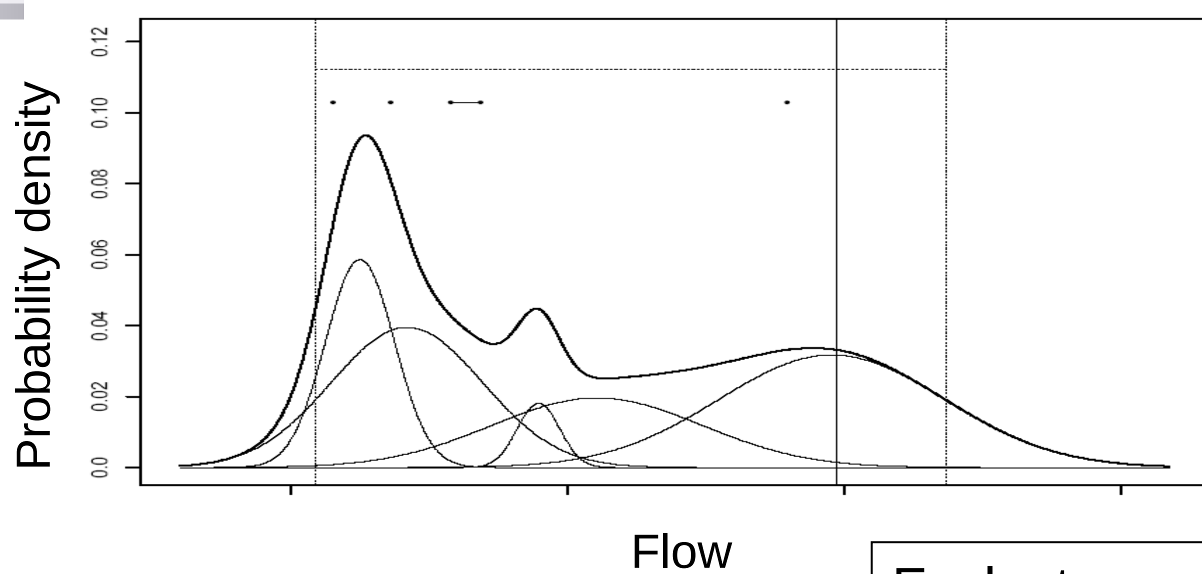
What would the pdf look like for a perfect forecast system?

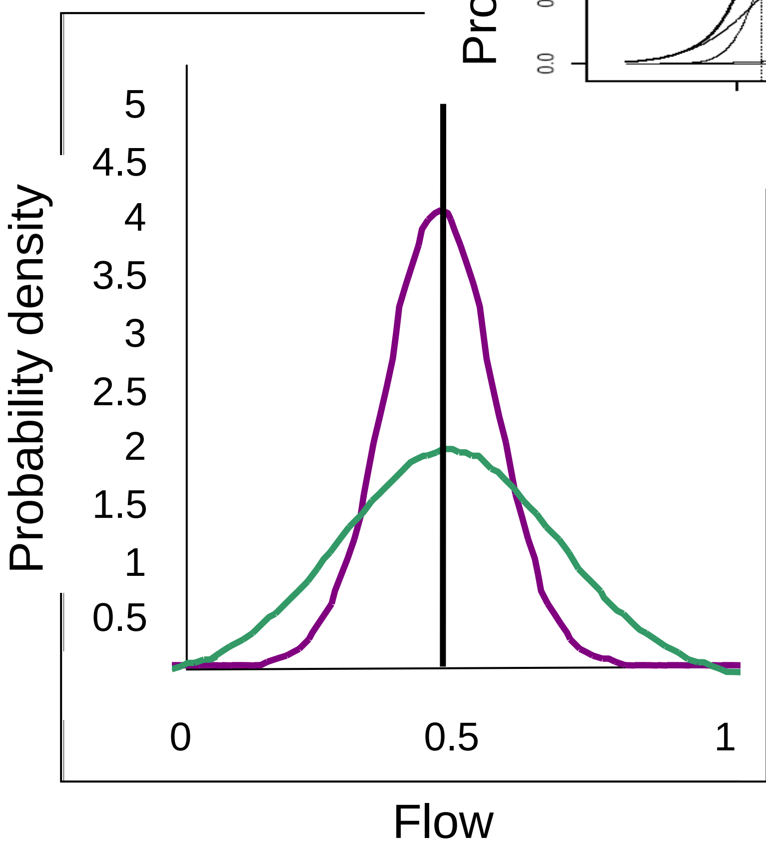
What is the central tendency for this forecast (e.g., ensemble mean)?

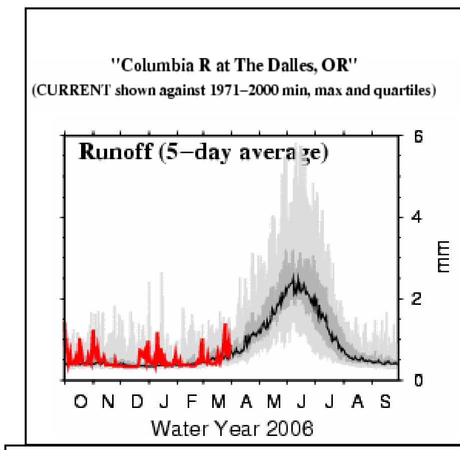
What do you expect the observed value will be?

How would you interpret this forecast for a user?

Comparing Deterministic & Ensemble Forecasts



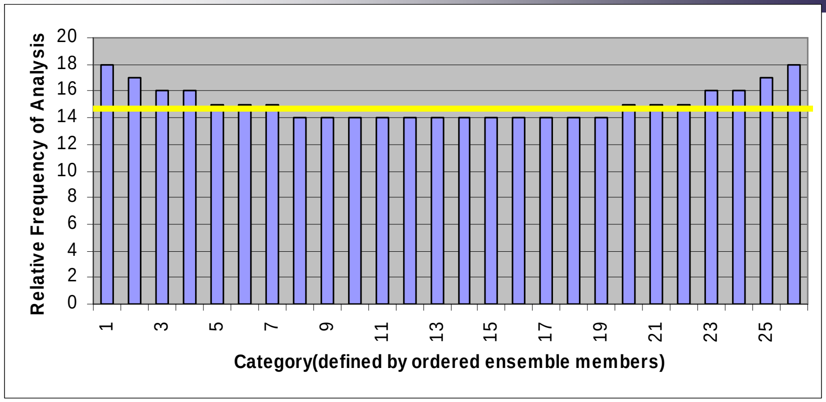




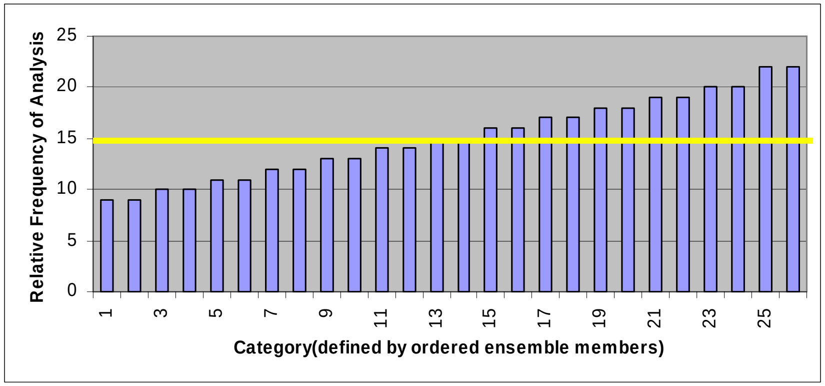
The mean minimizes error, but doesn't represent the overall behavior.

Evaluate deterministic forecasts using probabilistic metrics, rather than converting ensemble forecasts to single values (e.g., ensemble mean or median).

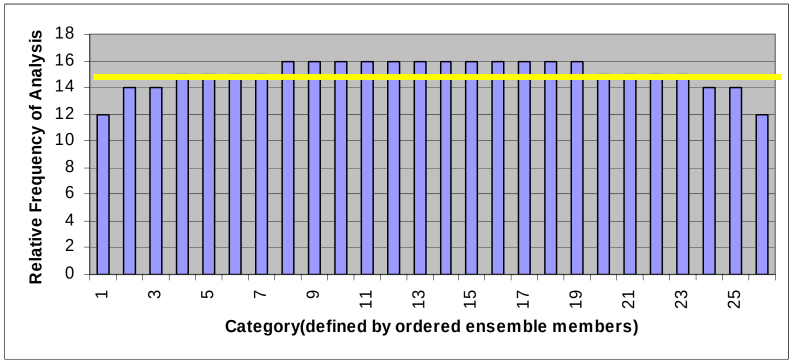
Rank Histogram: Needs lots of forecasts and observations



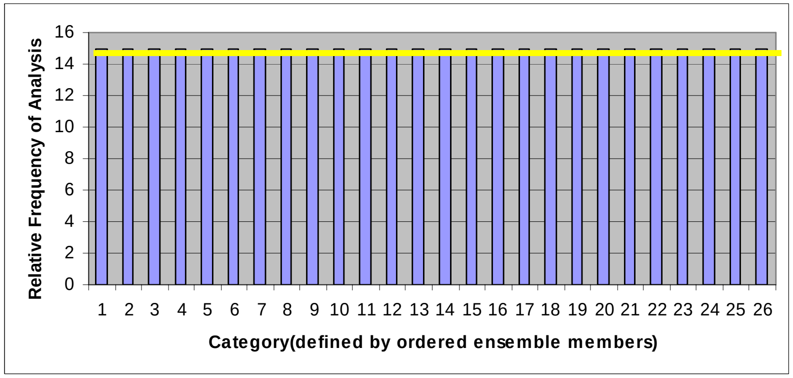
Example: “U-Shaped”
Indicates Ensemble Spread Too Small



Example: “L-Shaped”
Indicates Over or **Under** Forecasting Bias



Example: “N-Shaped” (domed shaped)
Indicates Ensemble Spread is Too Big



Example: “Flat-Shaped”
Indicates Ensemble Distribution Has Been Sampled Well

Courtesy K. Werner, NWS

When several categories are important…

INTRODUCTION TO
VERIFICATION
OF
HYDROLOGIC FORECASTS

3X3 Contingency Table

3 x 3 Contingency Table - Numerical

		Event Observed			
		Below 20 flow units	20-25 flow units	Above 25 flow units	Total
Event Forecasted	Below 20 flow units	a	b	c	a+b+c
	20-25 flow units	d	e	f	d+e+f
	Above 25 flow units	g	h	i	g+h+i
	Total	a+d+g	b+e+h	c+f+i	n

• Need for more than two verification categories

When would multiple categories be important?

Risk Management Perspective on Categories

Where all quantiles are based on prior analyses: flood stage, historical CDF, water rights, etc.

Two Category Forecast

$D_{\text{ecision}} = F\ (\ .75[I_{<\text{flood stage}}\ ] + .25[I_{=>\text{flood stage}}])$

Three Category Forecast

$\text{Decision} = F\ (\ .75[I_{<\text{flood stage}}] + .20[I_{=\text{flood}<\text{major flood stage}}] + .05[I_{=>\text{major flood}}])$

Risk Management Perspective on Categories

Where all quantiles are based on prior analyses: flood stage, historical CDF, water rights, etc.

Two Category Forecast

$D_{ecision} = F\ (\ .75[I_{<median}\ ] +\ .25[I_{=>median}])$

Three Category Forecast

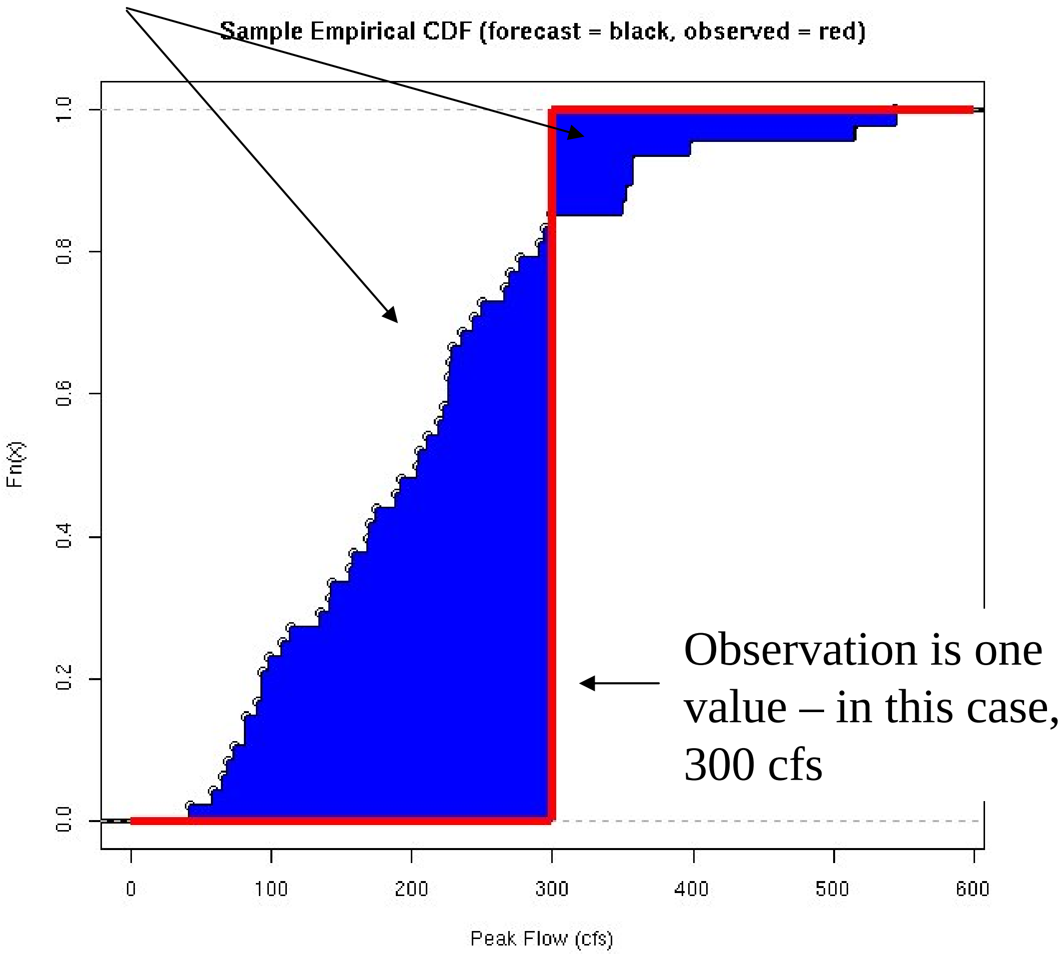
$Decision = F\ (\ .75[I_{quantile<10}] + .20[I_{quantile10\text{-}90}] + .05[I_{=>90}])$

- **Strategies: benefit from more categories, if sufficient skill**
- **Ideal: customized percentile categories**

Who decides which categories? How do they decide?

Continuous RPS Formulation: Many "Categories"

The RPS compares the forecast and observed cdfs.
Graphically, the CRPS is this area.



What would the empirical pdf look like for a perfect forecast system?

What audience would be most interested in Continuous RPS?

Courtesy Kevin Werner, NWS

Forecast Reliability

For a specified forecast condition, what does the distribution of observations look like?

P(O|F)

"When you say 80% chance of reaching flood stage, how often does flood stage occur?"

Reliability is conditioned on the forecasts (i.e., given that X was predicted, what was the outcome?). **The forecast probability is for a specific 'event', e.g., Peak Q<100cfs, Precip=>.25"**

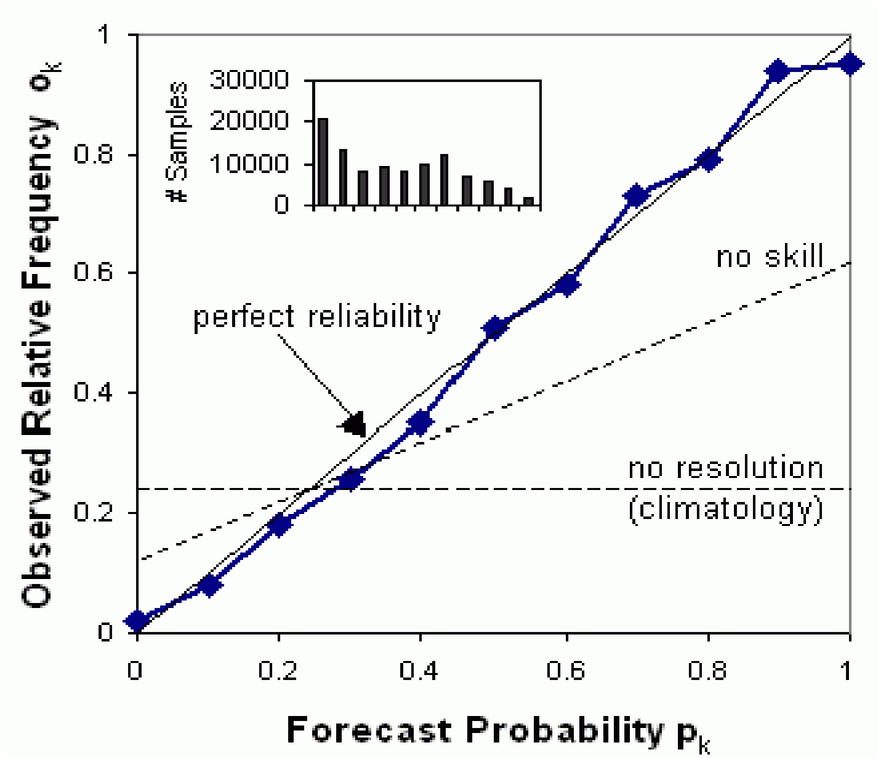
What audience would be interested in reliability?

How do you determine what 'event' to evaluate?

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.**

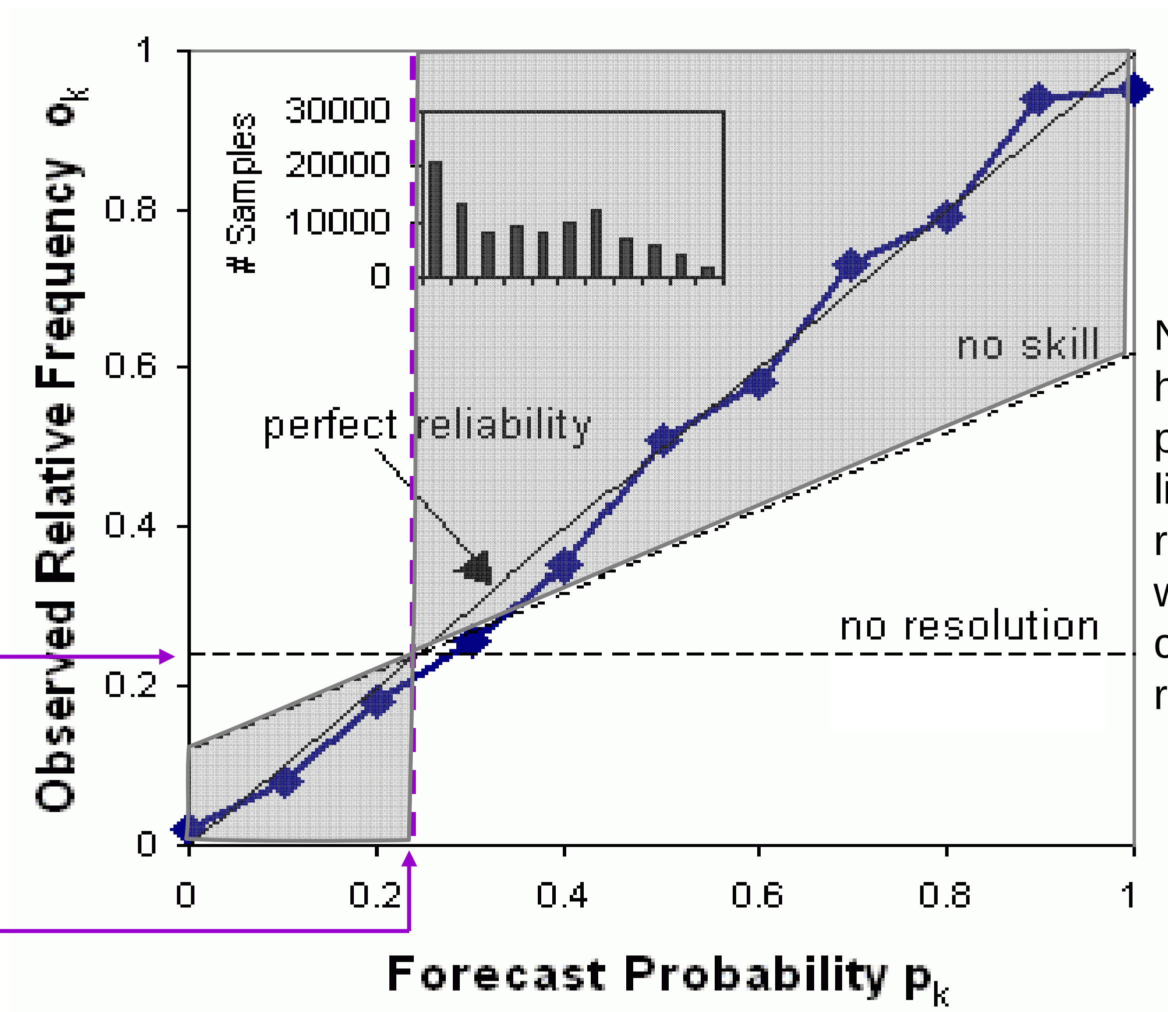


The reliability diagram is conditioned on the forecasts (i.e., given that X was predicted, what was the outcome?). **<u>The forecast probability is for a specific ‘event’, e.g., Peak Q<100cfs, Precip=>.25”</u>**

Source: Ebert (2005), courtesy of J. DeMargne

Attributes Diagram - Reliability, Resolution, Skill/No-skill

Overall relative frequency of observations (sample climatology)

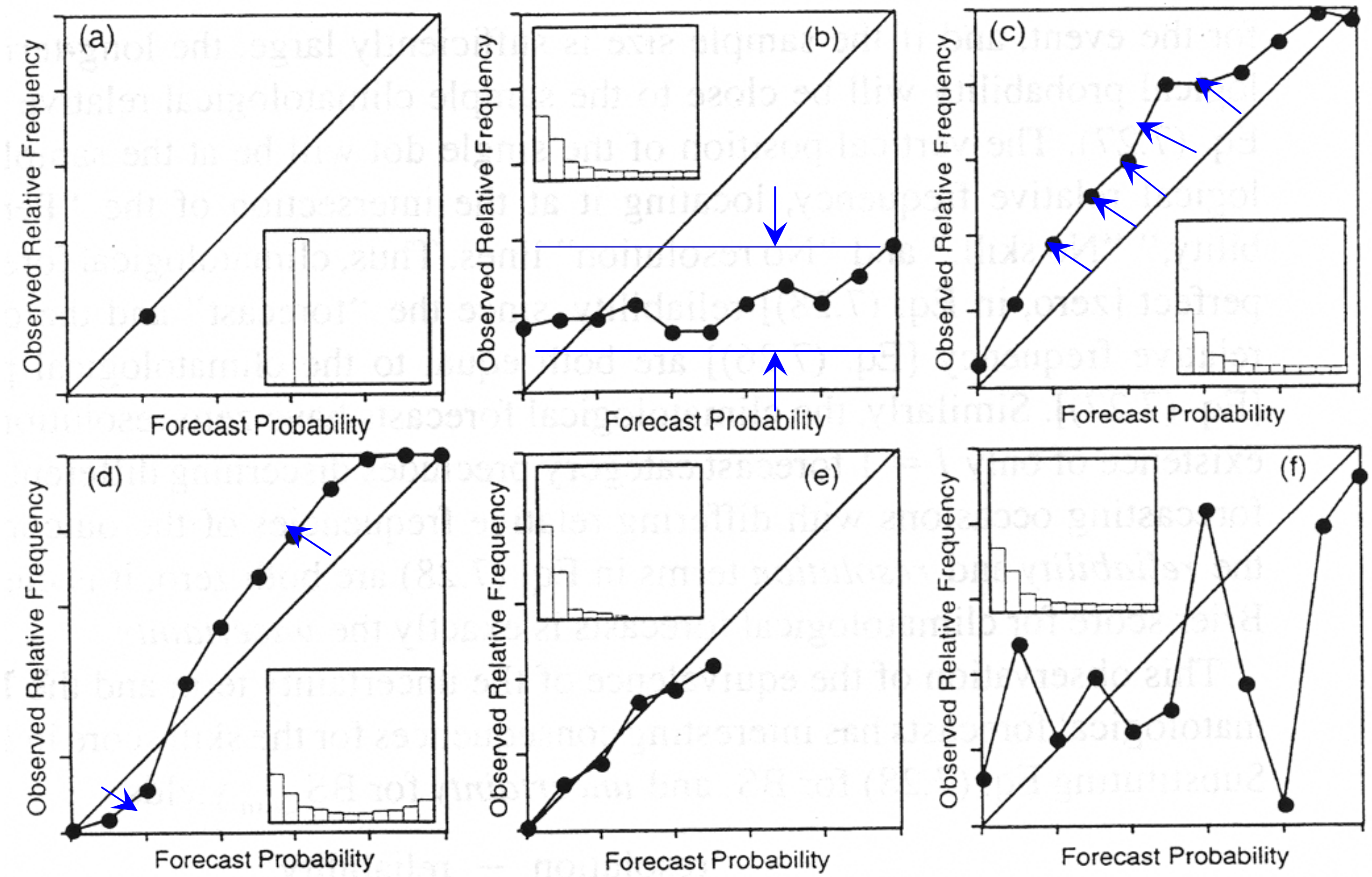


No-skill line : halfway between perfect-reliability line and no-resolution line, with sample climatology as a reference

Points closer to perfect-reliability line than to no-resolution line: subsamples of probabilistic forecast contribute positively to overall skill (as defined by BSS) in reference to sample climatology

Reliability: Attributes Diagram Interpretation

Climatology *Minimal RESolution* *Underforecasting*



Good RES, at expense of REL *Reliable forecasts of rare event* *Small sample size*

What does the reliability diagram look like for a perfect forecast system?

Source: Wilks (1995)

Forecast Discrimination

For a specified observation category, what do the forecast distributions look like?

P(F|O)

"When flood flows happened…
What were the forecasts saying?"

Discrimination is conditioned on the observations. When Y occurs, what do the forecast distributions look like? Do they look different than when X or Z occur?

Forecasts should look different when there's a flood, compared to when there's a drought!

What audience would be interested in discrimination?

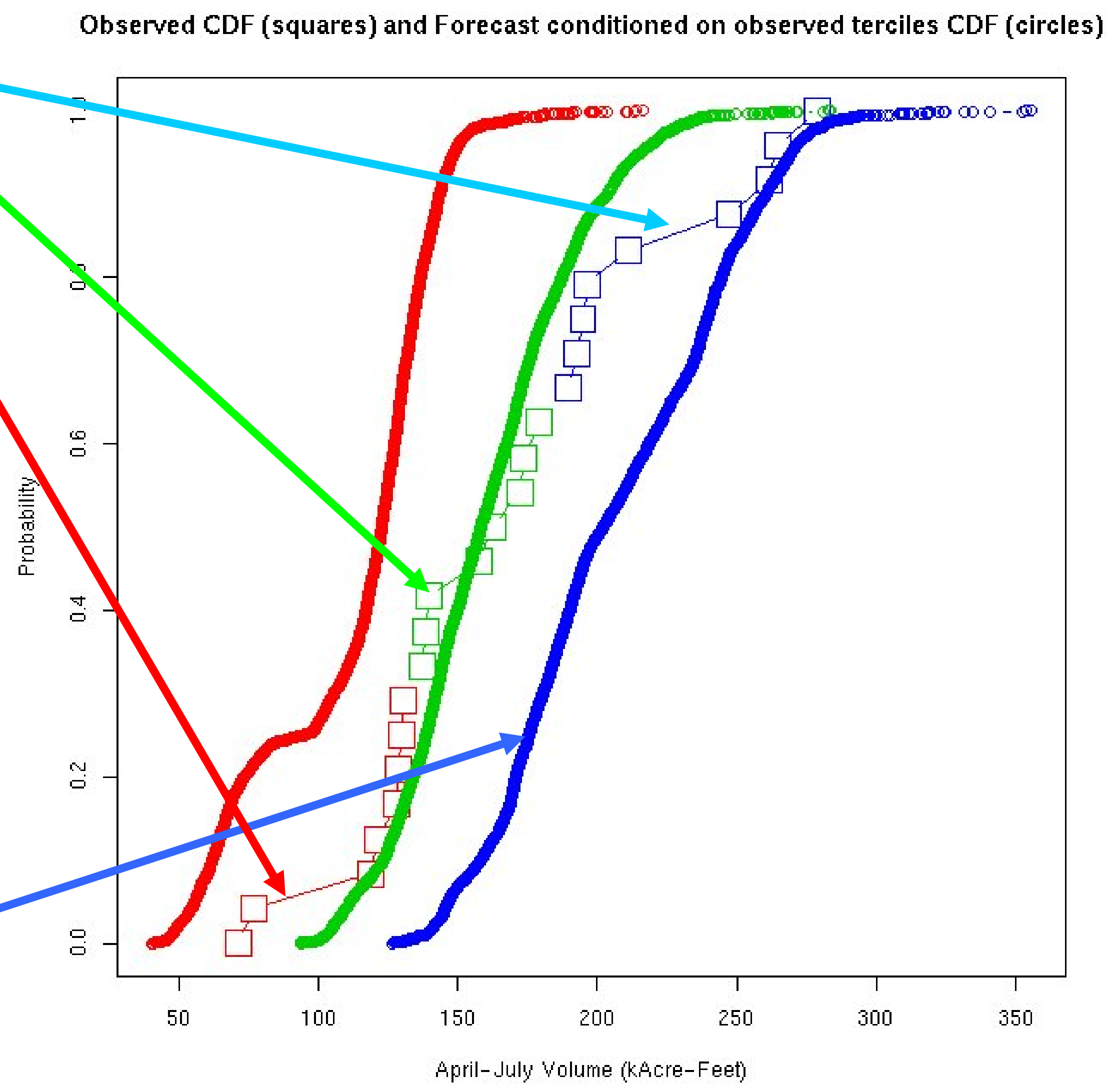
How do you determine what 'event' to evaluate?

Discrimination Example

All **observation** CDF is plotted and color coded by tercile.

Forecast ensemble members are sorted into 3 groups according to which tercile its associated observation falls into.

The CDF for each group is plotted in the appropriate color. i.e. high is blue.



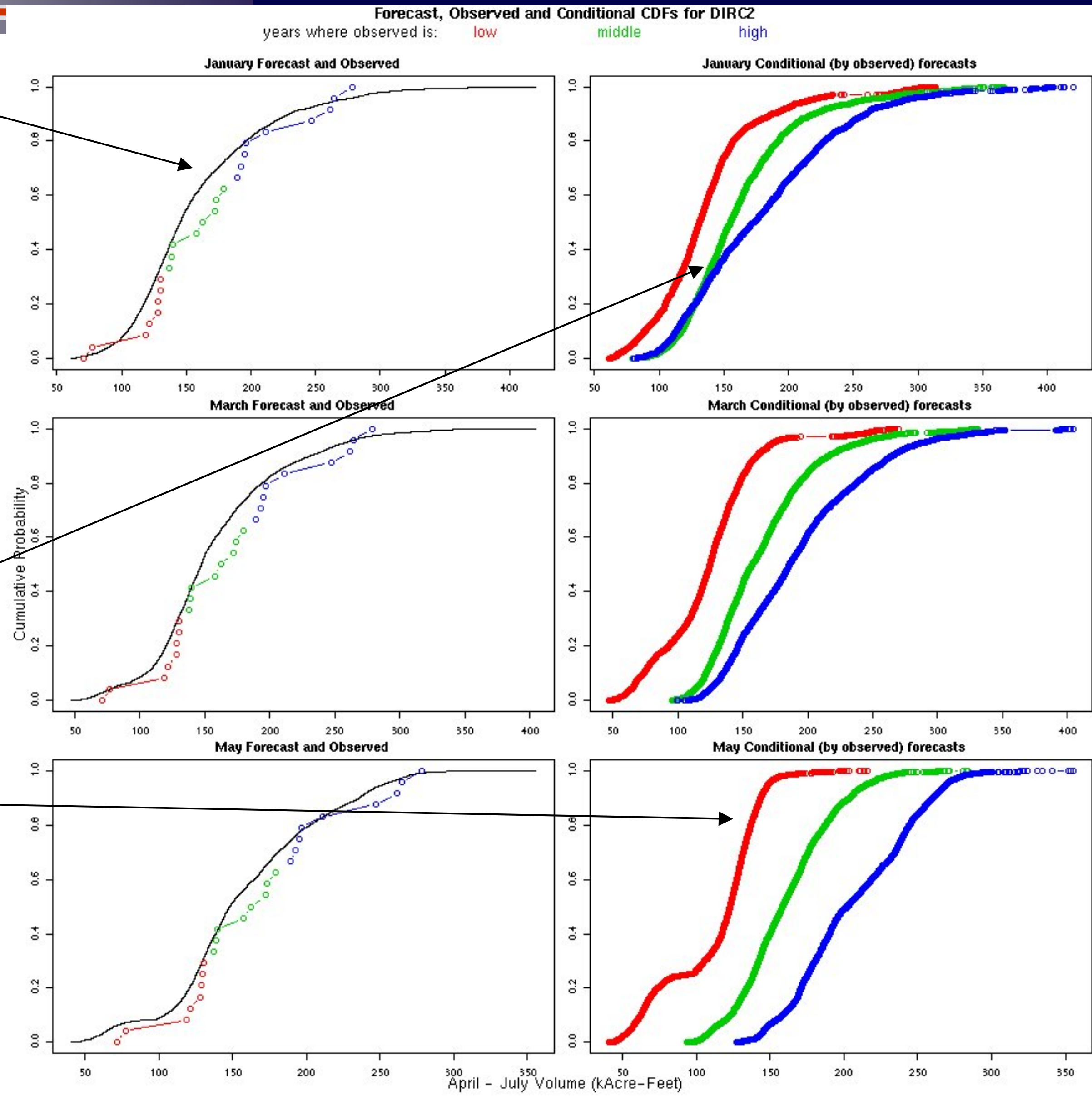
Courtesy K. Werner

Discrimination Example

Cdf for all forecast ensembles

How well do April – July volume forecasts discriminate when they are made in Jan, Mar, and May?

Poor discrimination in Jan between forecasting high and medium flows. Best discrimination in May.

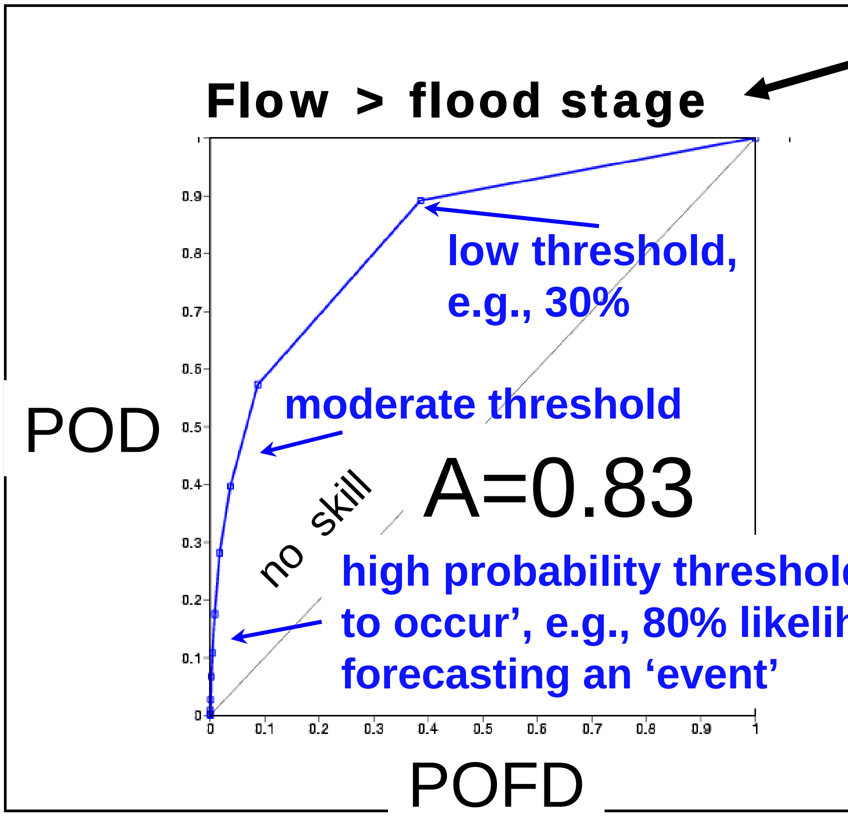


Relative Operating Characteristic (ROC)

ROC measures the ability of forecast to discriminate between events and non-events – <u>Conditioned on the Observations!</u>

ROC curve: plot of POD against POFD for range of probability thresholds

Conditional Event!



Source: Hagedorn (2006), courtesy J. Demargne

ROC area: area under the ROC curve; measures skill

A=0.5 => no skill

A=1 => perfect deterministic forecast

What would be a useful application of ROC?

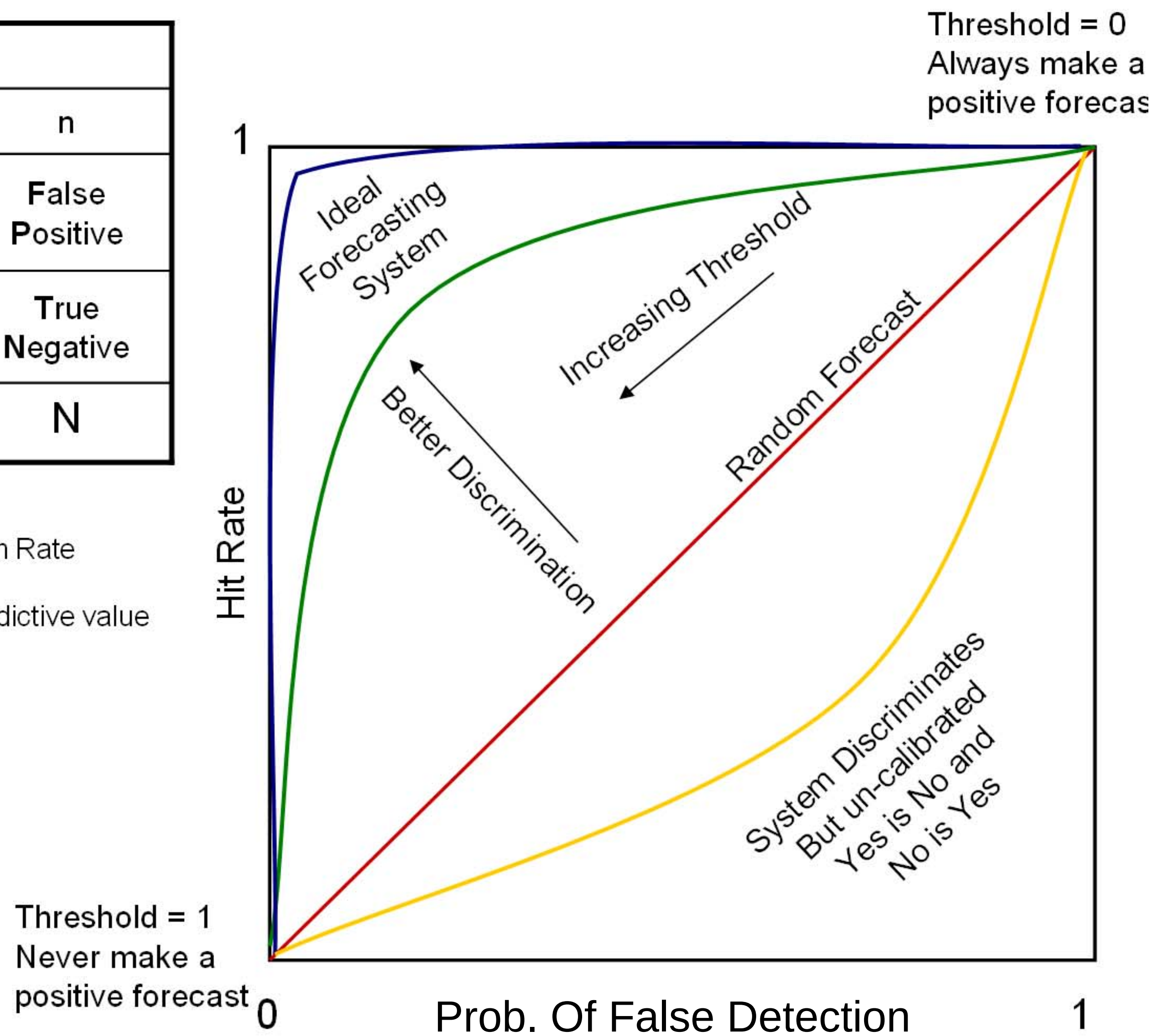
How would an emergency manager use the ROC?

Relative Operating Characteristic (ROC)

	Observation	
Forecast	p	n
Y	True Positive	False Positive
N	False Negative	True Negative
Total	P	N

False Positive rate=FP/N = False Alarm Rate
True Positive Rate= TP/P = Hit Rate
Precision = TP/(TP+FP) = Positive Predictive value
Accuracy= (TP+TN)/(P+N)
Specificity = TN/(FP+TN)

Note: Uses POD and POFD (Hit and Miss Rates), not FAR



Additional Topics in Verification

Sample Size

- Event assessment vs. forecast verification
- especially limiting for long-term forecasts, conditional measures, rare events
- Confidence bands on verification statistics, too!

Uncertainty in observations, too! Especially for major floods…

Timing Errors: Use time metrics, e.g., time of peak, time to drop below flood stage

Scheafli et al (2007)

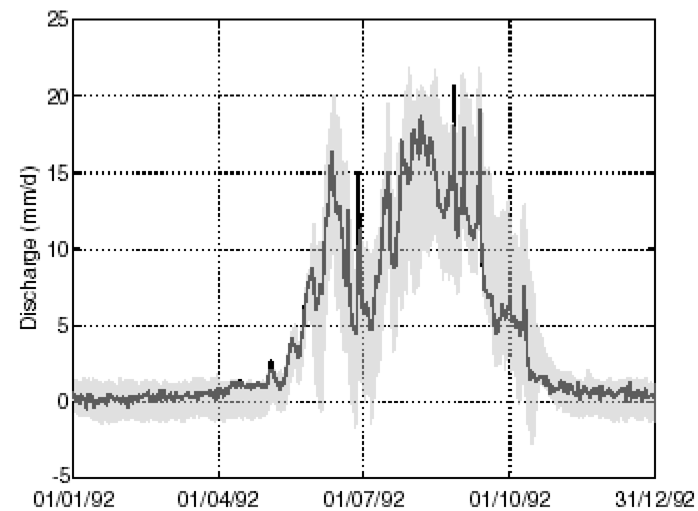


Figure 7 Observed discharge during validation period and 90% credibility interval induced by parameter uncertainty and modeling error (the negative observed values are due to the measurement error); for better readability only 1 year is shown.

Verification Strategies

Complete retrospective performance of ESP for all possible forecasts (Full Hindcast)

Skill of forecasts for the current forecast window, from previous years

Skill of recent forecasts leading up to the current forecast window

Evaluation of forecasts for periods having similar climatic and hydrologic conditions

What are the archive requirements to implement the full set of verification strategies?

Forecast Evaluation: Critical Needs

Multi-dimensional, distributions-oriented evaluation of all forecasts.

Compare by converting deterministic forecasts to probabilistic form – NOT the other way around.

Address small sample sizes for operational forecasts: Evaluate hindcasts for individual forecast techniques, objective forecast combinations, or pseudo-forecasts.

Incorporation of verification uncertainty! Confidence bounds on forecast verification as well as on the forecasts themselves.

Consider uncertainty in observations in verification.
Better estimation of naturalized flows.

Communication of forecast performance to users.
Cooperation of forecasting agencies and external groups.