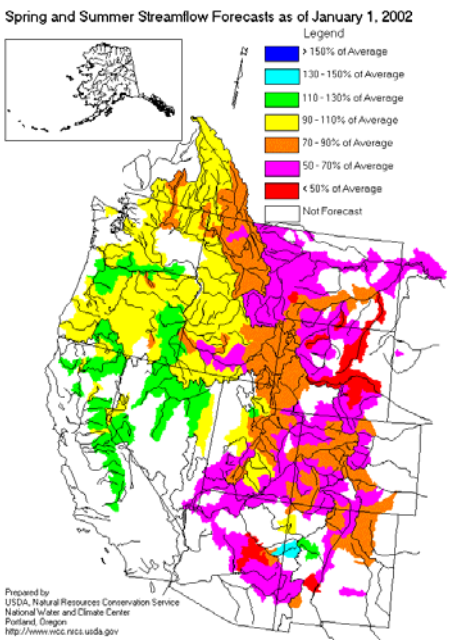
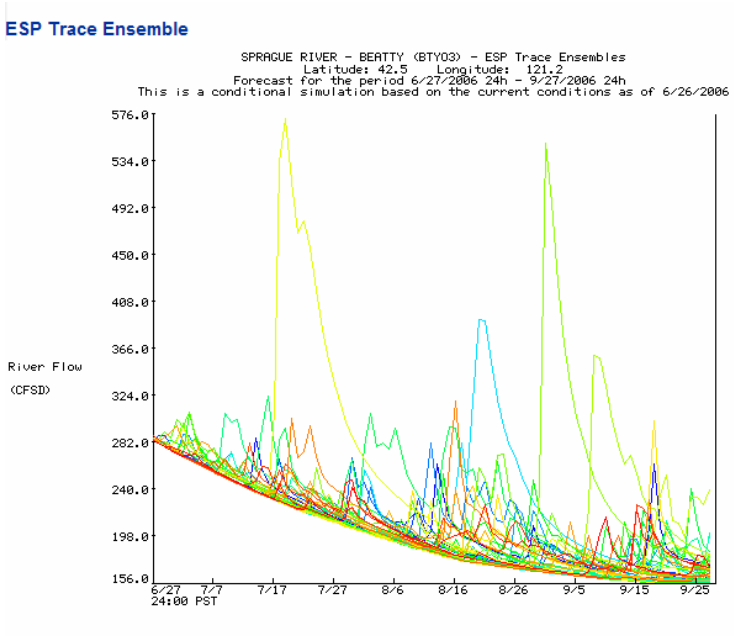
Supported by:

Hydrologic Forecast Verification





Holly C. Hartmann

Department of Hydrology and Water Resources, University of Arizona

hollyoregon@juno.com

NOAA GAPP

NOAA CLIMAS

HyDIS: NASA/Raytheon Synergy

NSF SAHRA

NWS CSD

Goals

- General concepts of verification
- Think about how to apply to your operations
- Be able to respond to and influence NWS verification program
- Be prepared as new tools become available
- Be able to do some of their own verification
- Be able to work with researchers on verification projects
- Contribute to development of verification tools (e.g., look at various options)
- Avoid some typical mistakes

Why Do Verification?

Administrative: logistics, selected quantitative criteria

Operations: inputs, model states, outputs, quick!

Research: sources of error, targeting research

Users: making decisions, exploit skill, avoid mistakes

Concerns about verification?

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

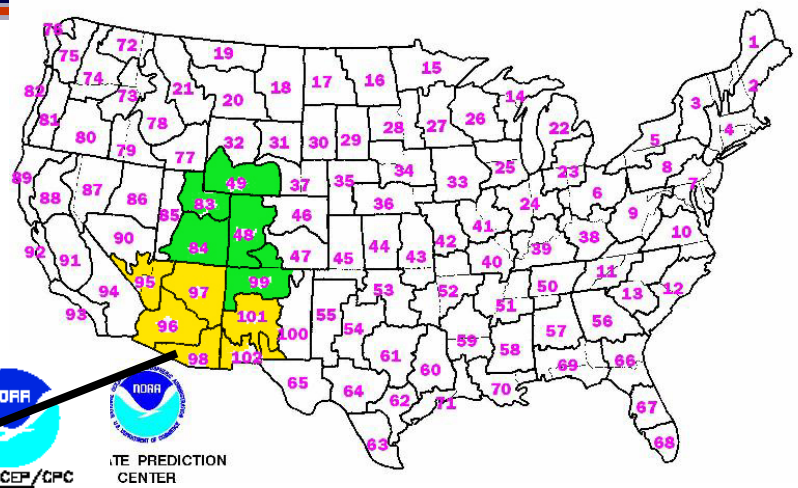
Relevant forecast variables, regions (location & scale), seasons, lead times, performance characteristics

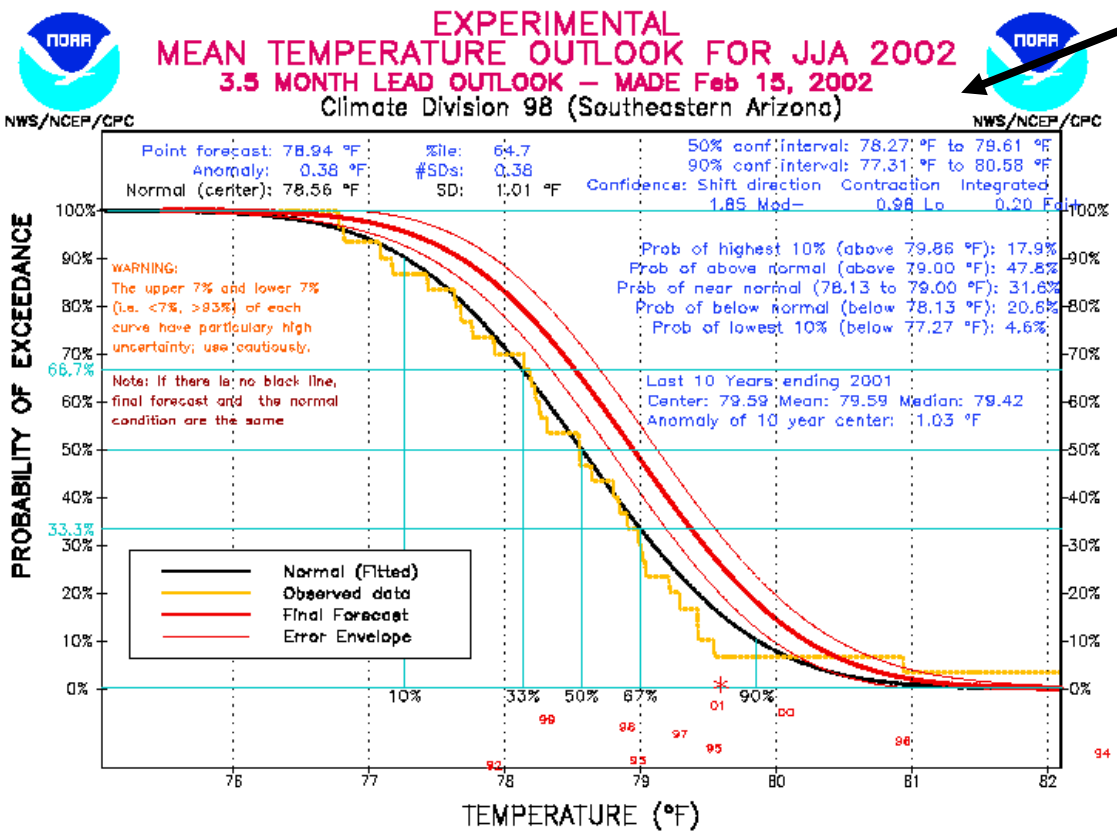
Technical sophistication: base probabilities, distributions, math

Role of of forecasts in decision making

Prob. Forecasts: User preferences influence verification

Probability of Exceedance Forecasts: These forecasts say something about the entire range of possibilities (not just at tercile boundaries). They provide probabilities and quantities for individual locations.





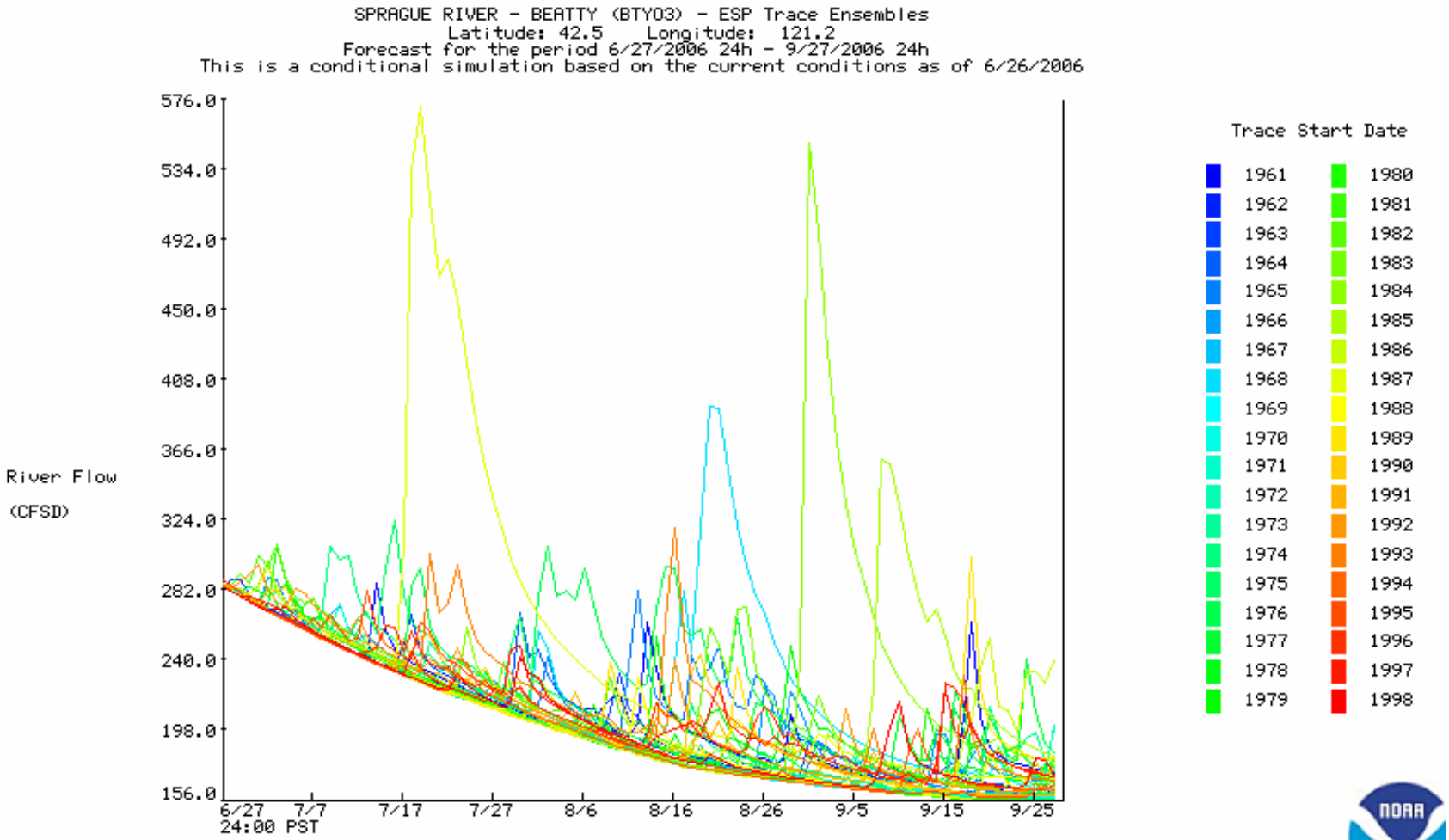
Although these forecasts are more difficult to understand, they contain <u>much</u> more information than any of the previously available forecast formats.

They allow customized forecasts via tradeoffs between 'confidence' and 'precision'.

www.cpc.ncep.noaa.gov/pacdir/NFORdir/

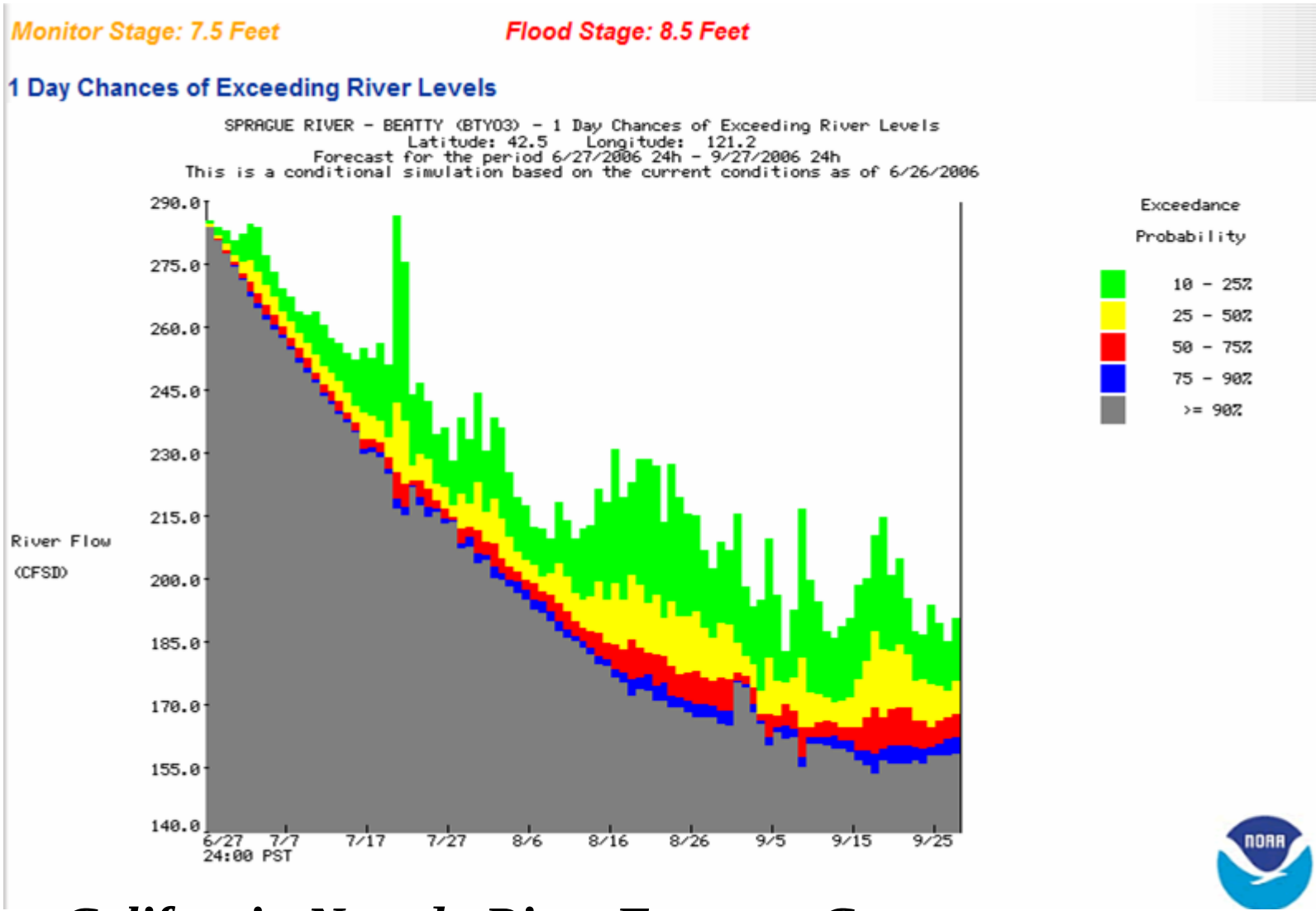
ESP Forecasts: User preferences influence verification

ESP Trace Ensemble



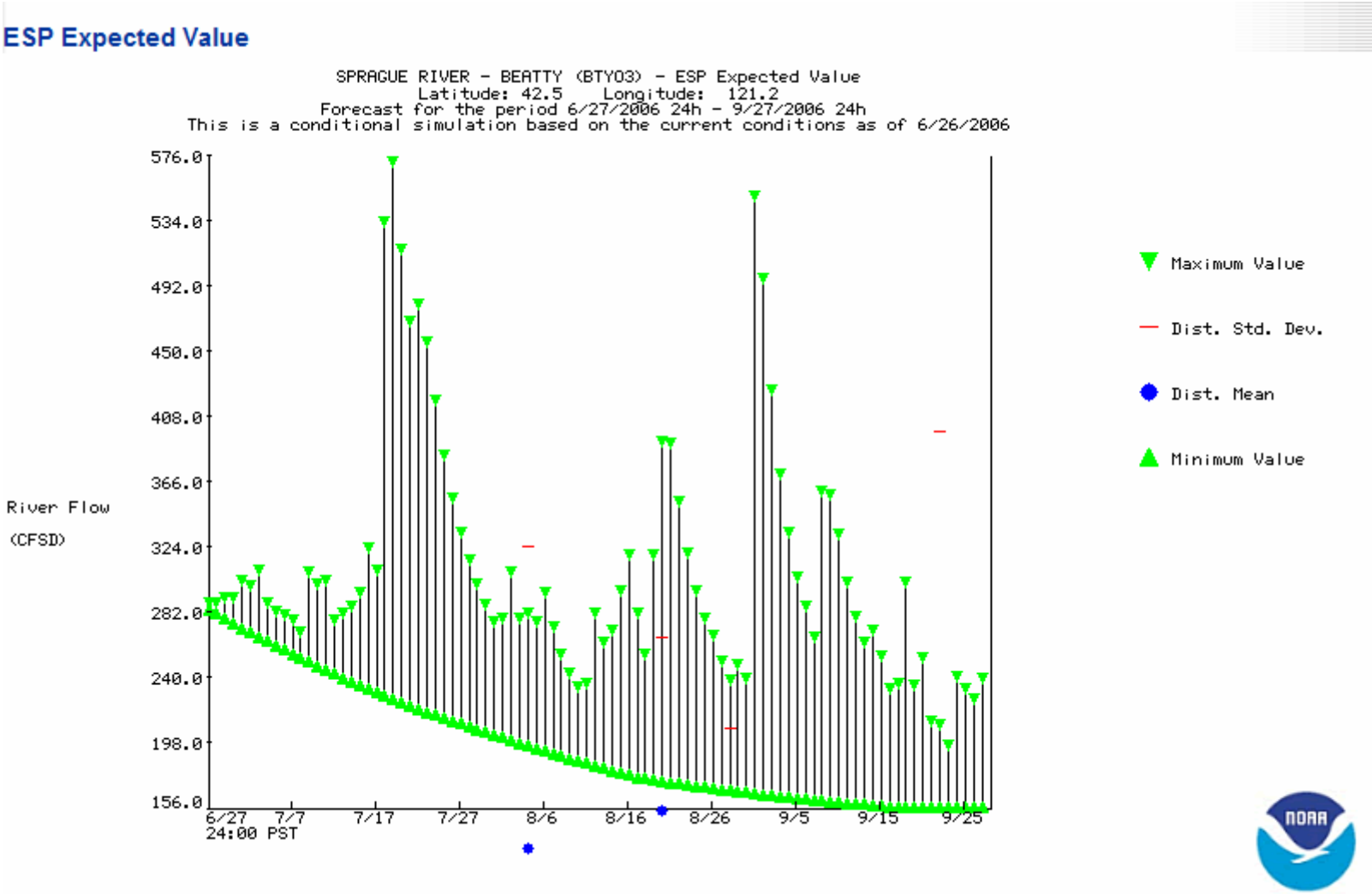
From: California-Nevada River Forecast Center

ESP Forecasts: User preferences influence verification



From: California-Nevada River Forecast Center

ESP Forecasts: User preferences influence verification

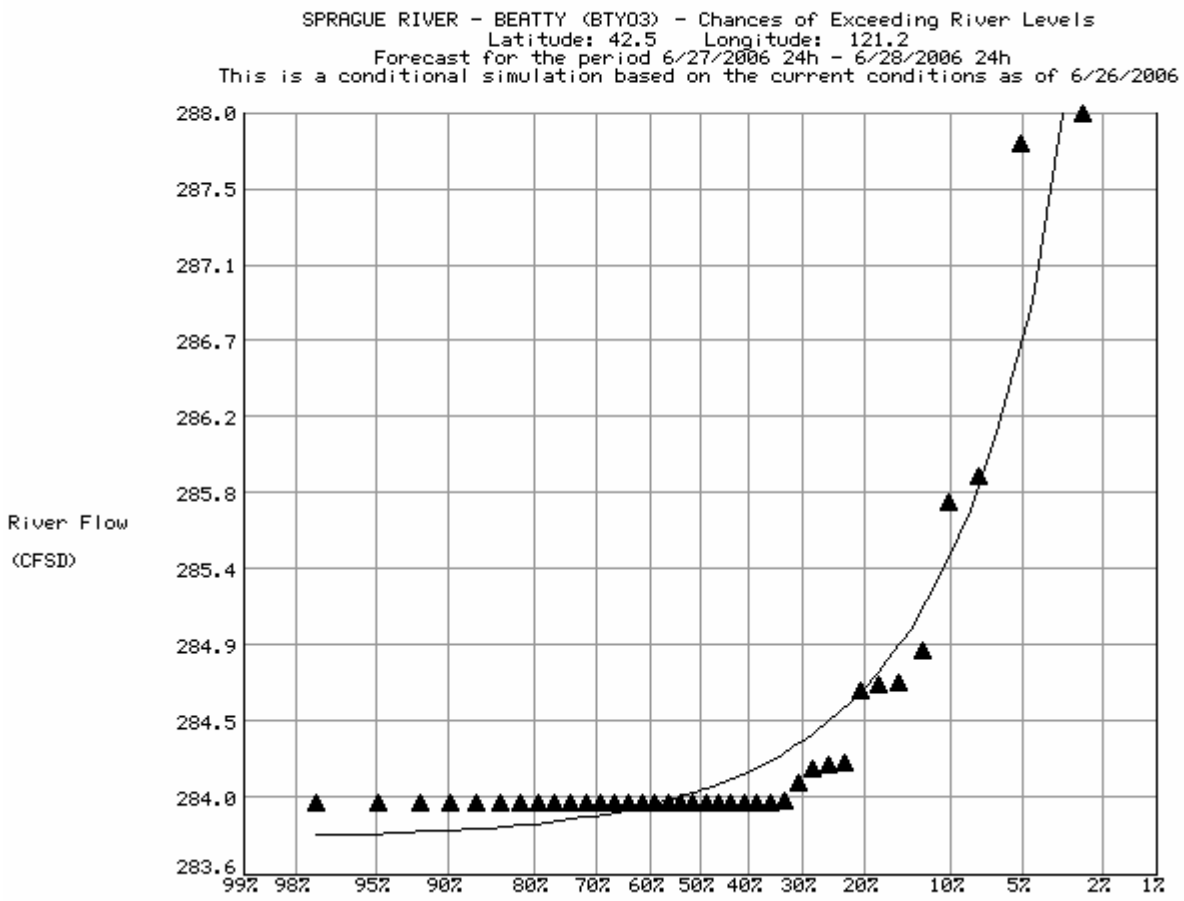


From: California-Nevada River Forecast Center

ESP Forecasts: User preferences influence verification

Monitor Stage: 7.5 Feet *Flood Stage: 8.5 Feet*

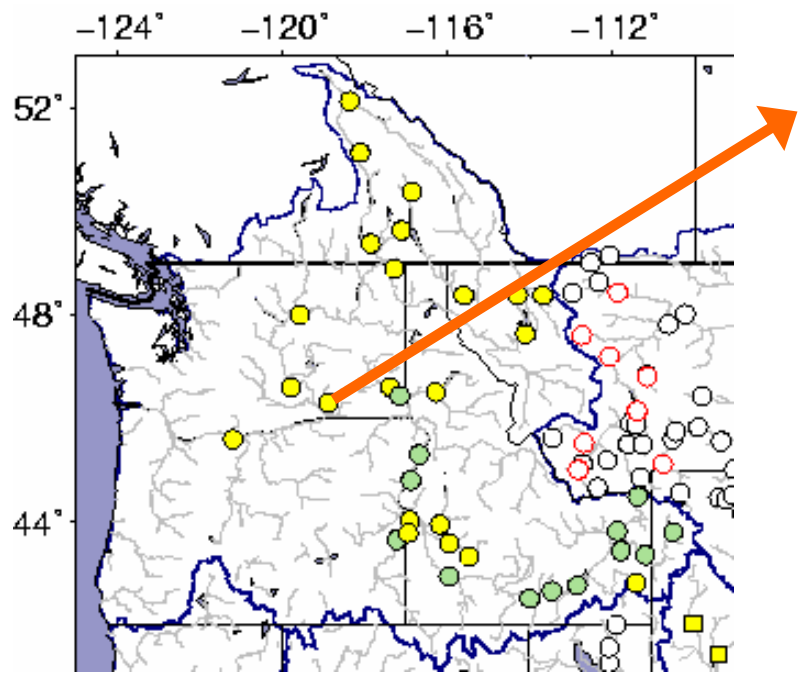
Chances of Exceeding River Levels

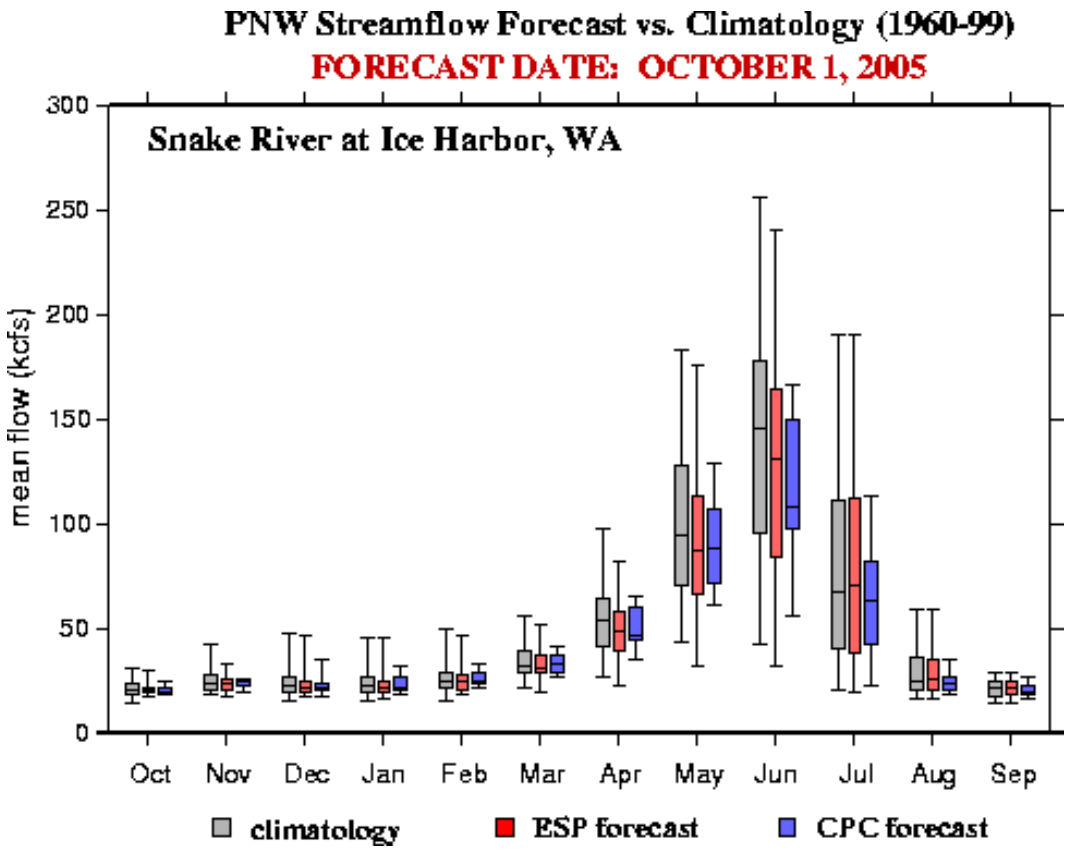


From: California-Nevada River Forecast Center

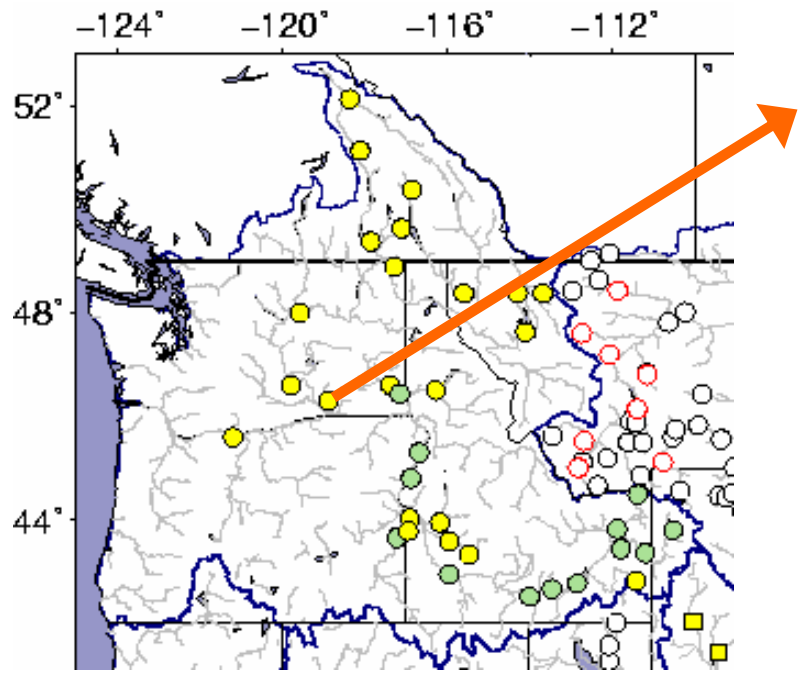
Probabilistic ESP Forecasts

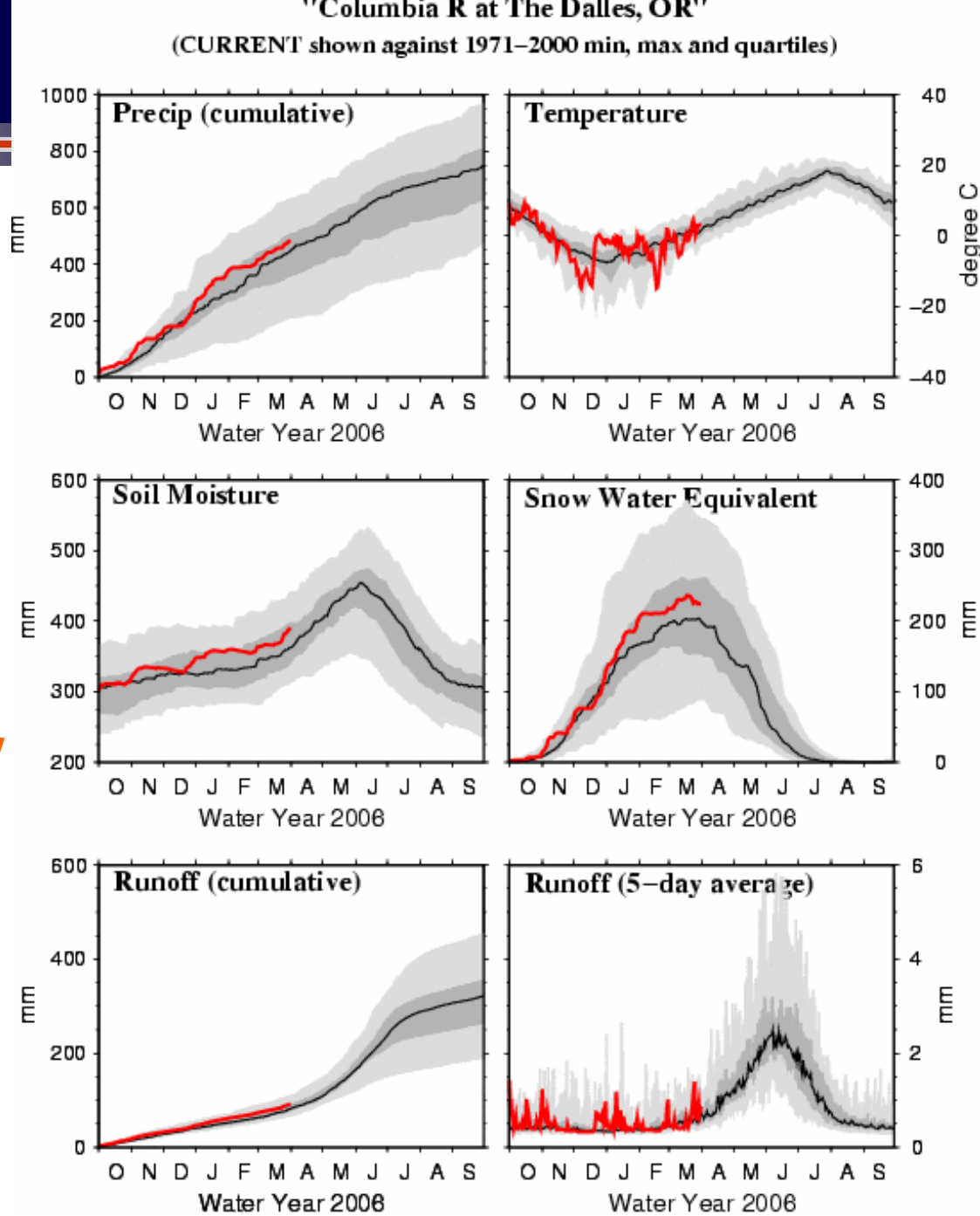
From: A. Hamlet, University of Washington





From: A. Hamlet,
University of Washington





Different Forecasts, Information, Evaluation

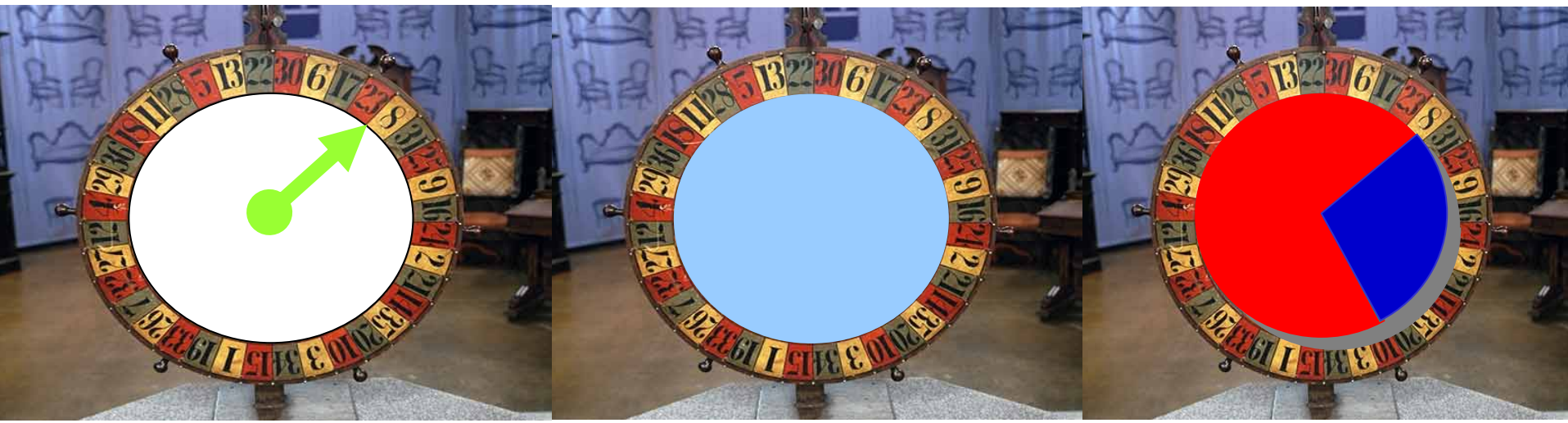
"Today's high will be 76 degrees, and it will be partly cloudy, with a 30% chance of rain."

Deterministic

Categorical

Probabilistic

Deterministic ***Categorical*** ***Probabilistic***



How would you evaluate each of these?

So Many Evaluation Criteria!

Deterministic

Bias

Correlation

RMSE

• Standardized RMSE

• Nash-Sutcliffe

Linear Error in Probability Space

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
Percent N-class errors
Modified Heidke Skill Score
Hannsen and Kuipers Score
Gandin and Murphy Skill Scores…

Probabilistic

Brier Score

Ranked Probability Score

Distributions-oriented Measures

• Reliability

• Discrimination

• Sharpness

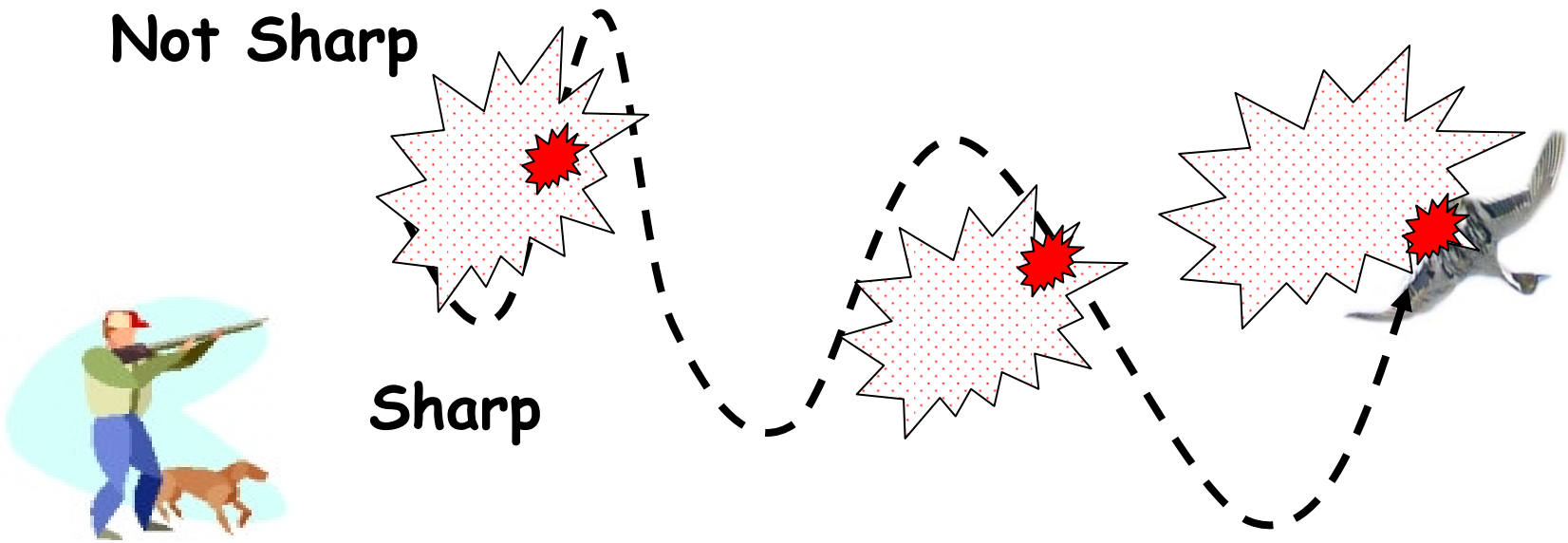
Possible Performance Criteria

Accuracy - *overall correspondence between forecasts and observations*

Bias - *difference between average forecast and average observation*

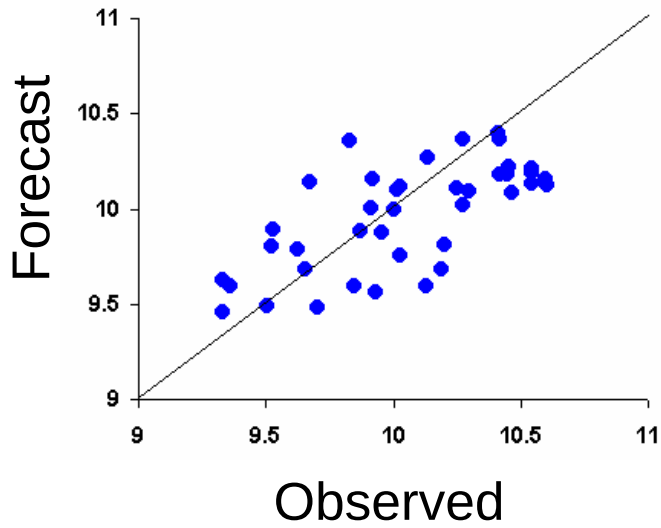
Consistency - *forecasts don't waffle around*

Sharpness/Refinement – *ability to make bullish forecast statements*



Bias

Mean forecast = Mean observed



Correlation Coefficient

Variance shared between forecast and observed

Says nothing about bias or whether

forecast variance = observed variance

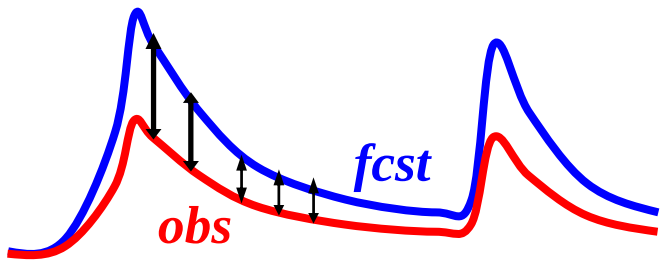
Root Mean Squared (Standard) Error

Distance between forecast/observation values

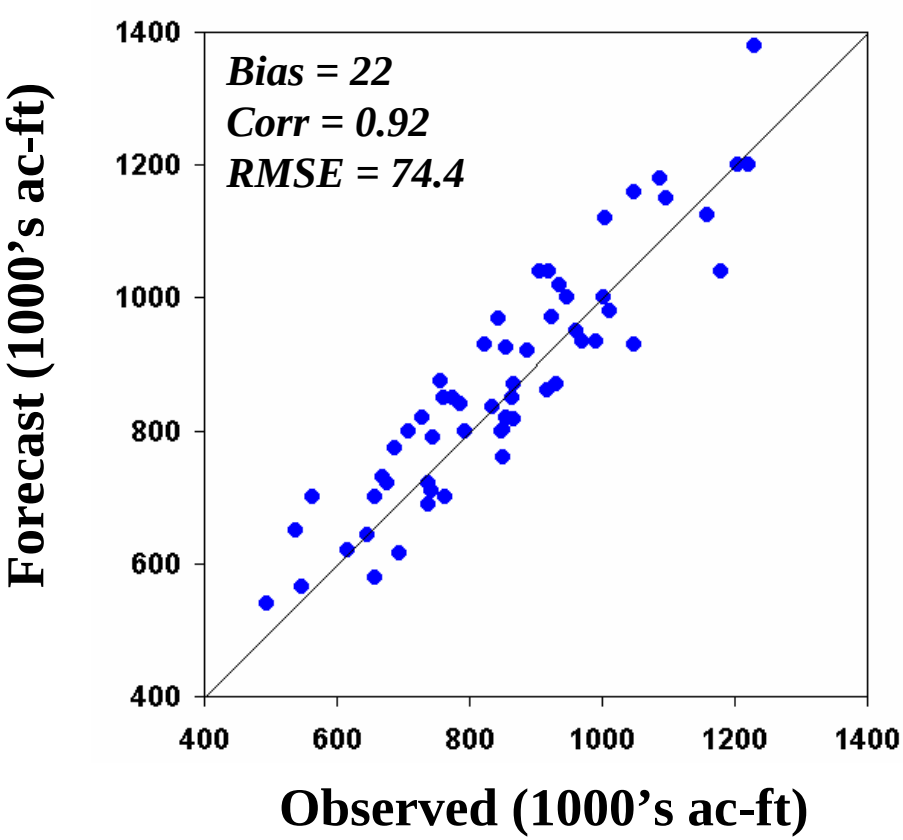
Better than correlation, but does poor when error is heteroscedastic

Emphasizes performance for high flows

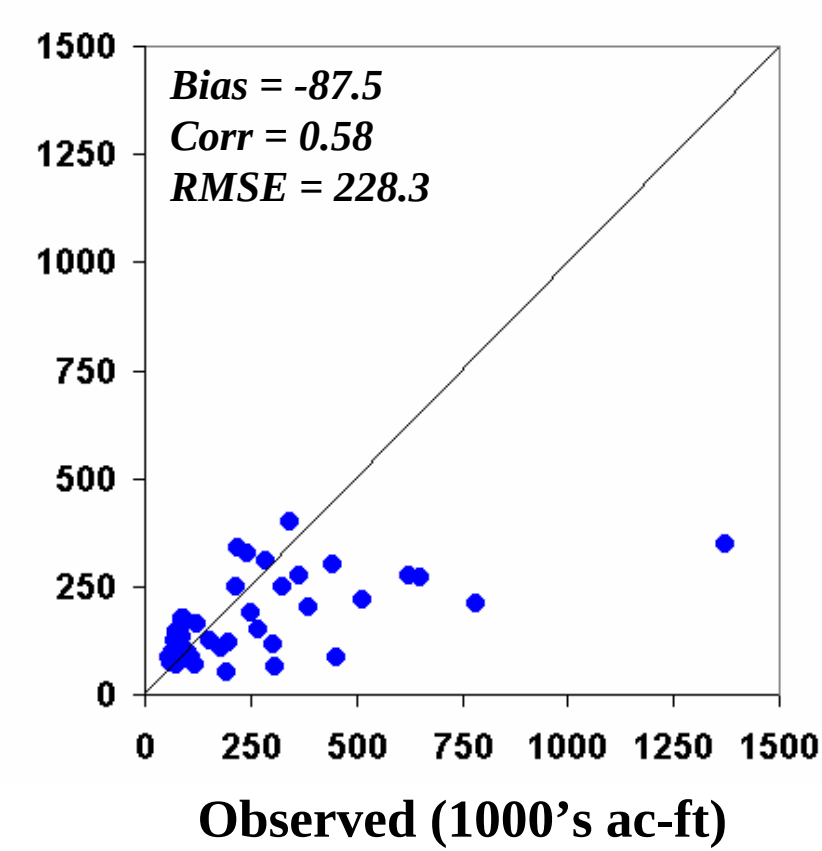
Alternative: Mean Absolute Error (MAE)



1943-99 April 1 Forecasts for Apr-Sept Streamflow at Stehekin R at Stehekin, WA



1954-97 January 1 Forecasts for Jan-May Streamflow at Verde R blw Tangle Crk, AZ



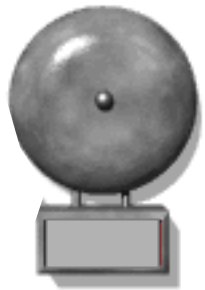
Forecasting Tradeoffs

Forecast performance is multi-faceted

False Alarms
warning without event

No fire

Surprises
event without warning



“False Alarm Rate”

“Probability of Detection”

A forecaster’s fundamental challenge is balancing these two.
Which is more important?
Depends on the specific decision context…

Contingency Table Evaluations: Ignore Probabilities

Flood Forecast?	Flood Observed? Yes	No	Total
Yes	10	20	30
No	35	35	70
Total	45	55	100

User Perspective: Only one category is relevant

Example: Flood forecast

Probability of detection: 10/45 = 22%

How often were you not 'surprised'?

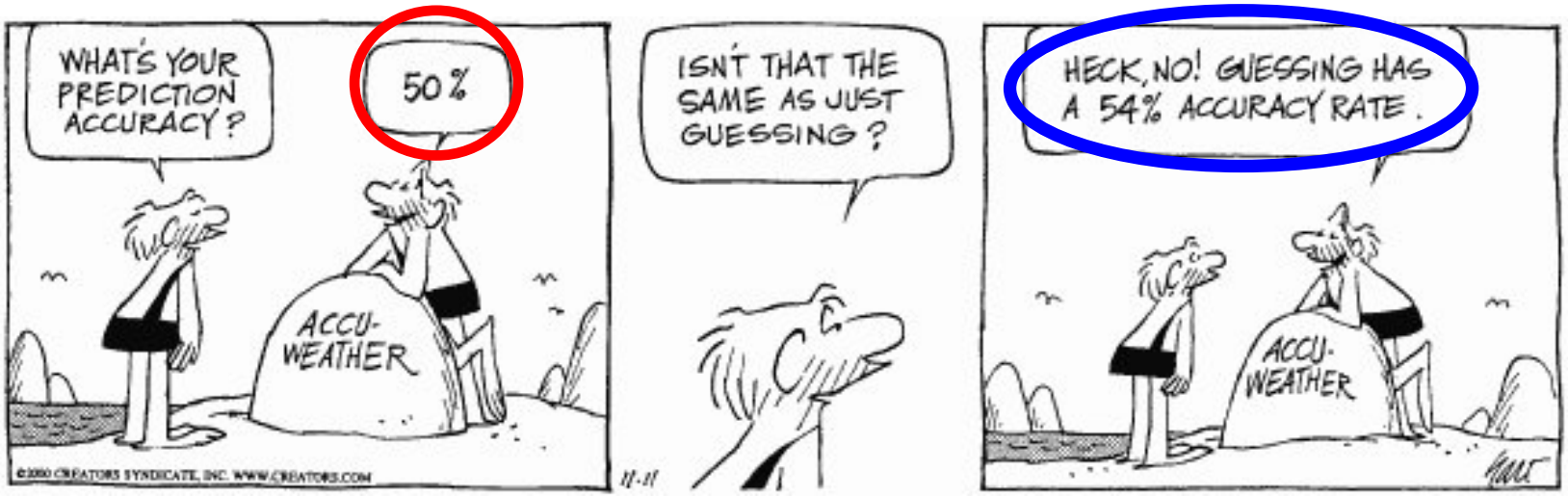
False Alarm Rate: 20/30 = 66%

How often were you 'led astray'?

But what did you expect by chance alone?

How Good? Compared to What?

$$\text{Skill Score} = \frac{\text{Forecast - Baseline}}{\text{Perfect - Baseline}}$$



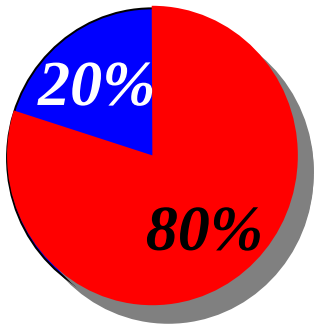
Skill: (0.50 – 0.54)/(1.00-0.54) = -8.6%

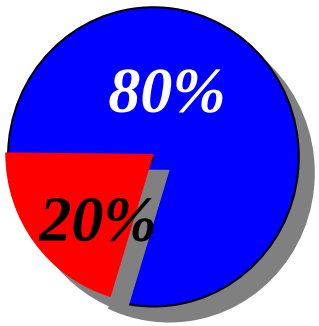
~worse than guessing~

What is the appropriate Baseline?

Probabilistic Forecast Evaluation: "Brier" Score

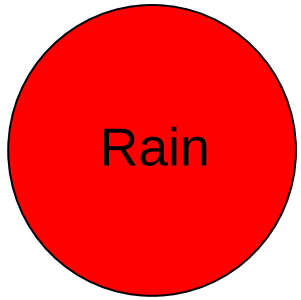
Forecast:
"80% chance of rain"





Climatology
(Baseline chances)

Observed Outcome



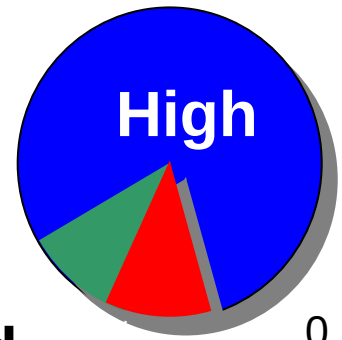
Good

Not Good

With this forecast, what outcome would you prefer?

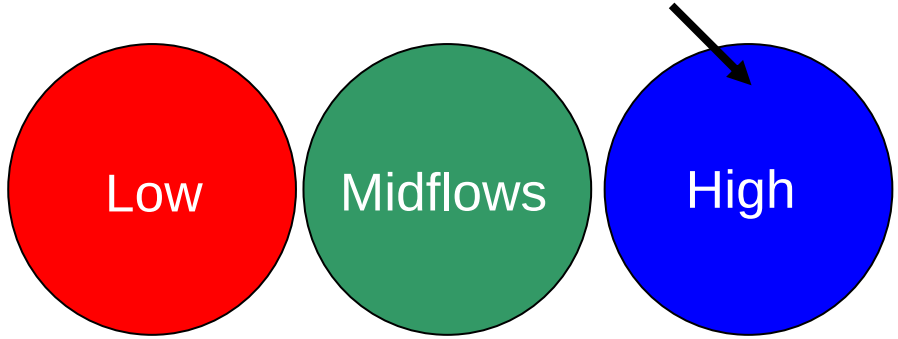
Probabilistic Evaluation: Ranked Probability Score

Bold Forecaster

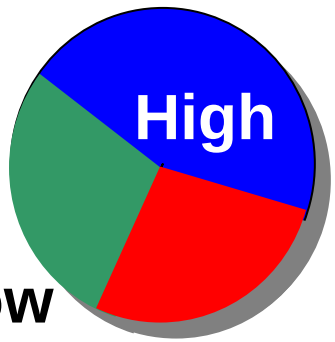


0.80/0.17/0.03

Observed Outcome



Conservative Forecaster



0.40/0.33/0.27

$RPSB = (0.03 - 0)^2 + (0.20 - 0)^2 + (1 - 1)^2 = 0.04$

$RPSC = (0.27 - 0)^2 + (0.60 - 0)^2 + (1 - 1)^2 = 0.43$

$RPS_{clim} = (0.30 - 0)^2 + (0.70 - 0)^2 + (1 - 1)^2 = 0.58$

$SSB_{rps} = (0.04 - 0.58)/(0 - 0.58) = 0.931 = 93\%$

$SSC_{rps} = (0.43 - 0.58)/(0 - 0.58) = 0.259 = 26\%$

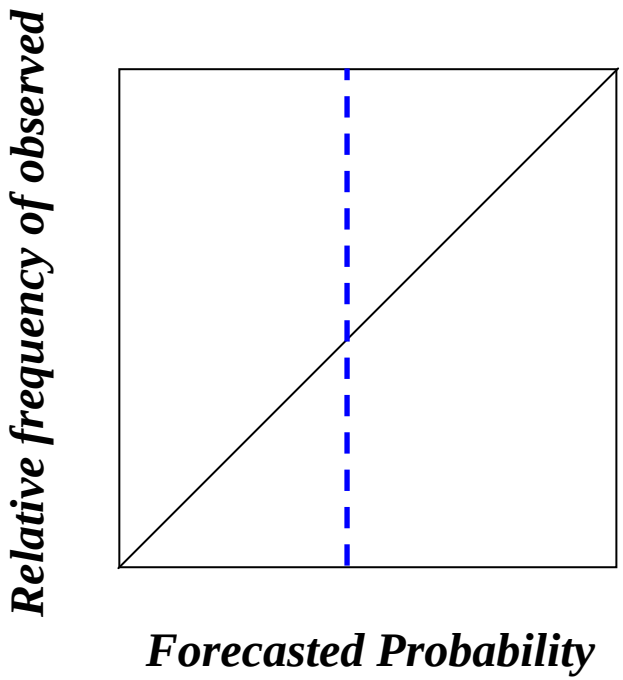
Reliability Diagrams

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

$$P(O|F)$$

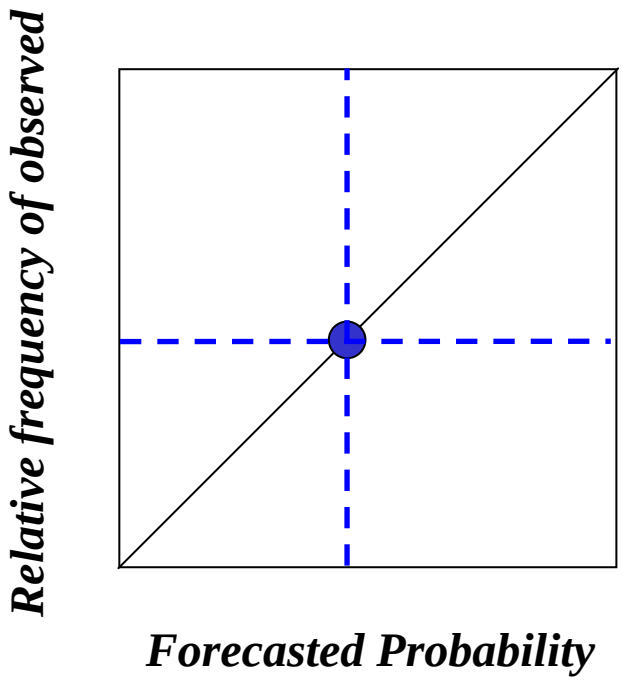
Forecast Reliability

If the forecast says
there's a 50% chance of high flows…



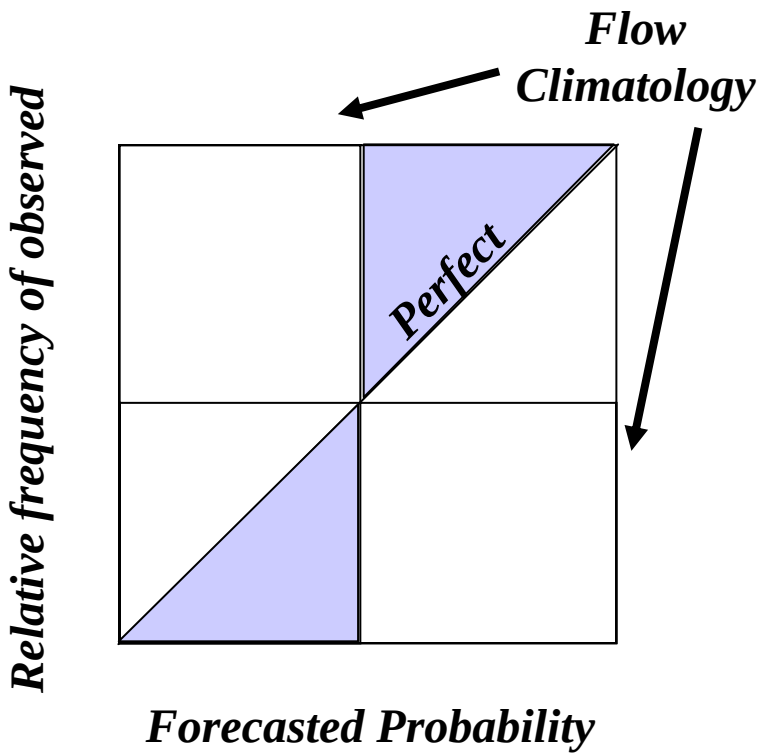
Forecast Reliability

If the forecast says
there's a 50% chance of high flows…
High flows should happen 50% of the time



Forecast Reliability

If the forecast says
there's a 50% chance of high flows…
High flows should happen 50% of the time



Flow "climatology": Median value

Forecasts "better" than expected. Probabilities could have been more extreme and maintained quality.

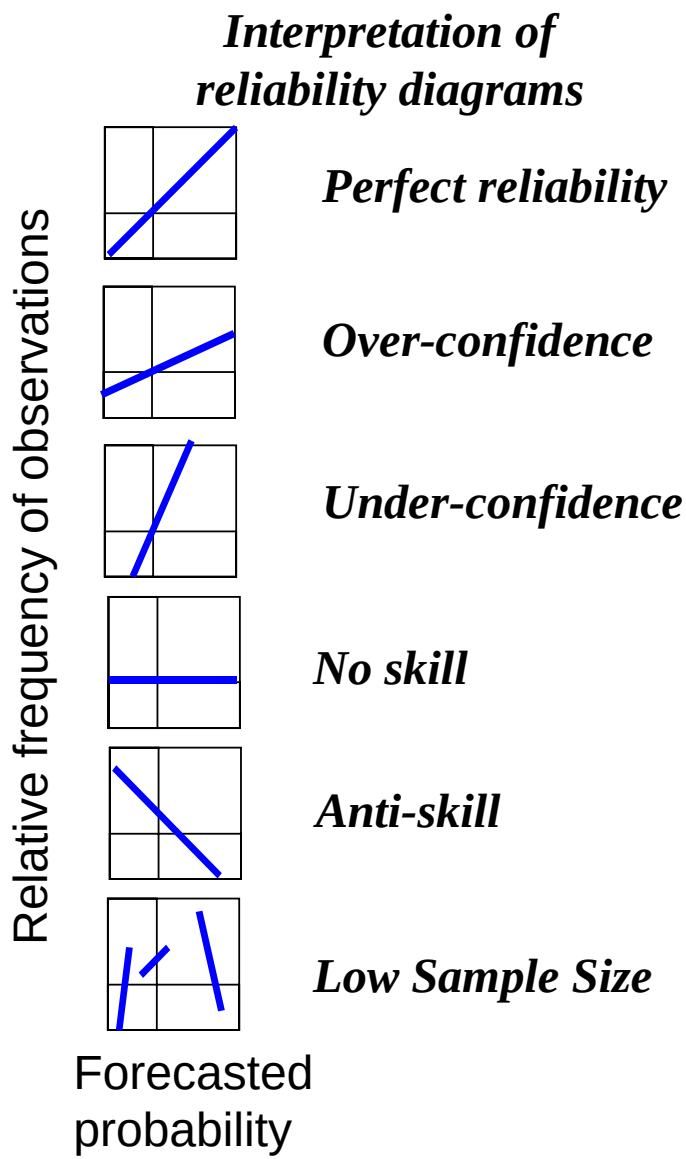
Interpretation of Reliability Diagrams

Reliability

P[O|F]

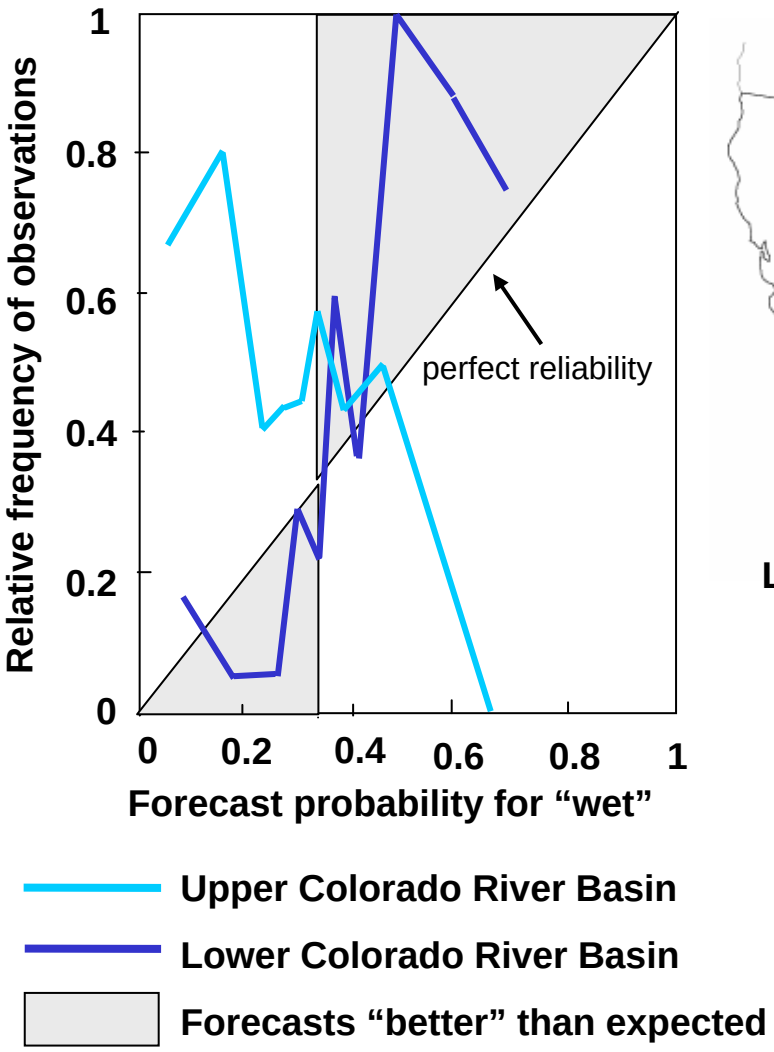
Does the frequency of occurrence match your probability statement?

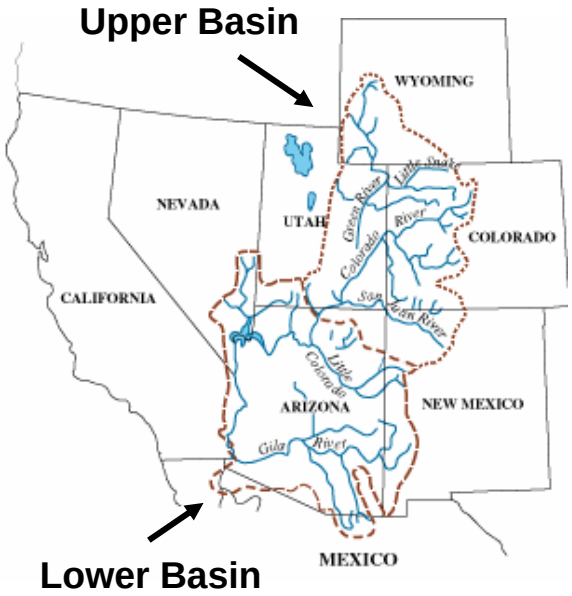
Identifies conditional bias



Reliability: CPC forecasts & water management

Precipitation forecasts accurately reflect expected performance



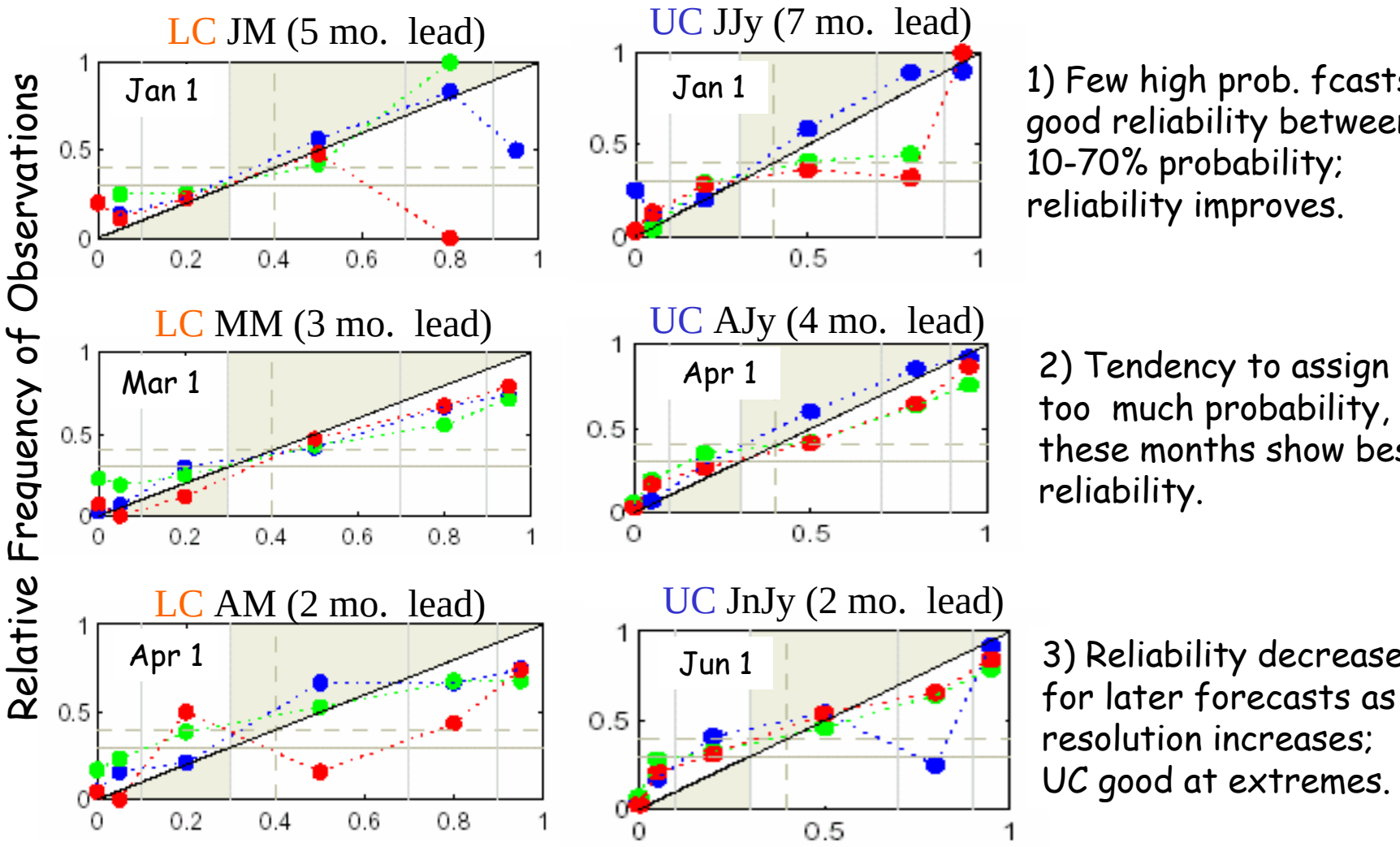


~1995-2001 winter season, summer/fall outlooks

CPC forecast performance varies among regions, with important implications for resource management.

Seasonal climate forecasts have been much better for the Lower Colorado Basin than for the Upper Basin.

Reliability: Colorado Basin ESP Seasonal Supply Outlooks



1) Few high prob. fcasts, good reliability between 10-70% probability; reliability improves.

2) Tendency to assign too much probability, these months show best reliability.

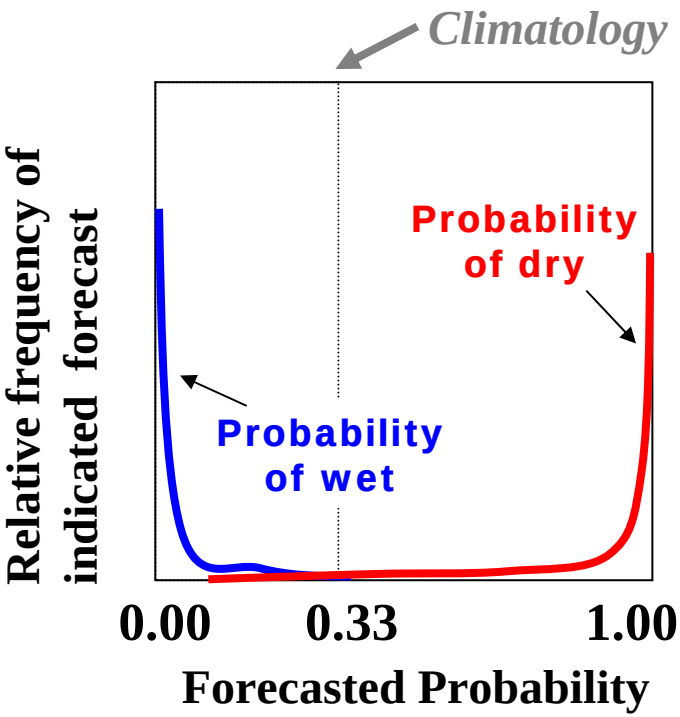
3) Reliability decreases for later forecasts as resolution increases; UC good at extremes.

Discrimination: CPC Climate Outlooks

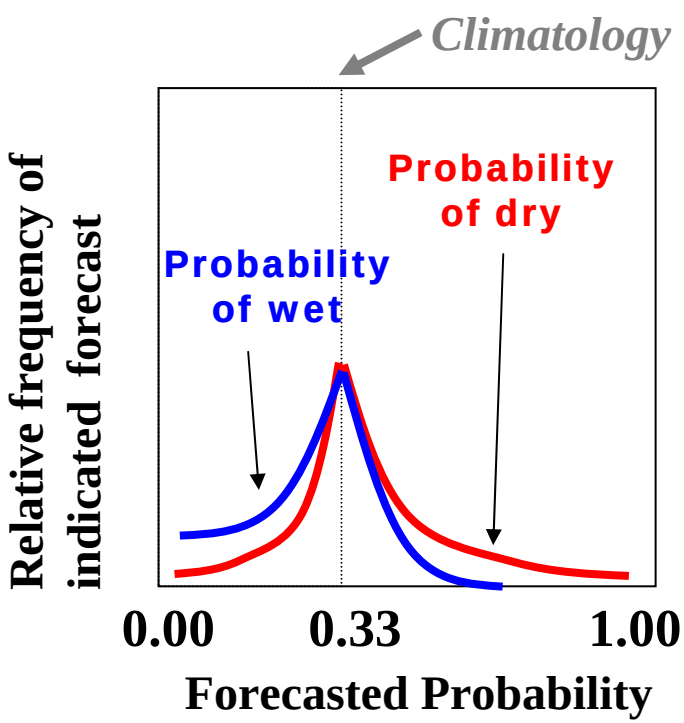
Discrimination: P[F|O]

Can the forecasts distinguish among different events?

Good discrimination!



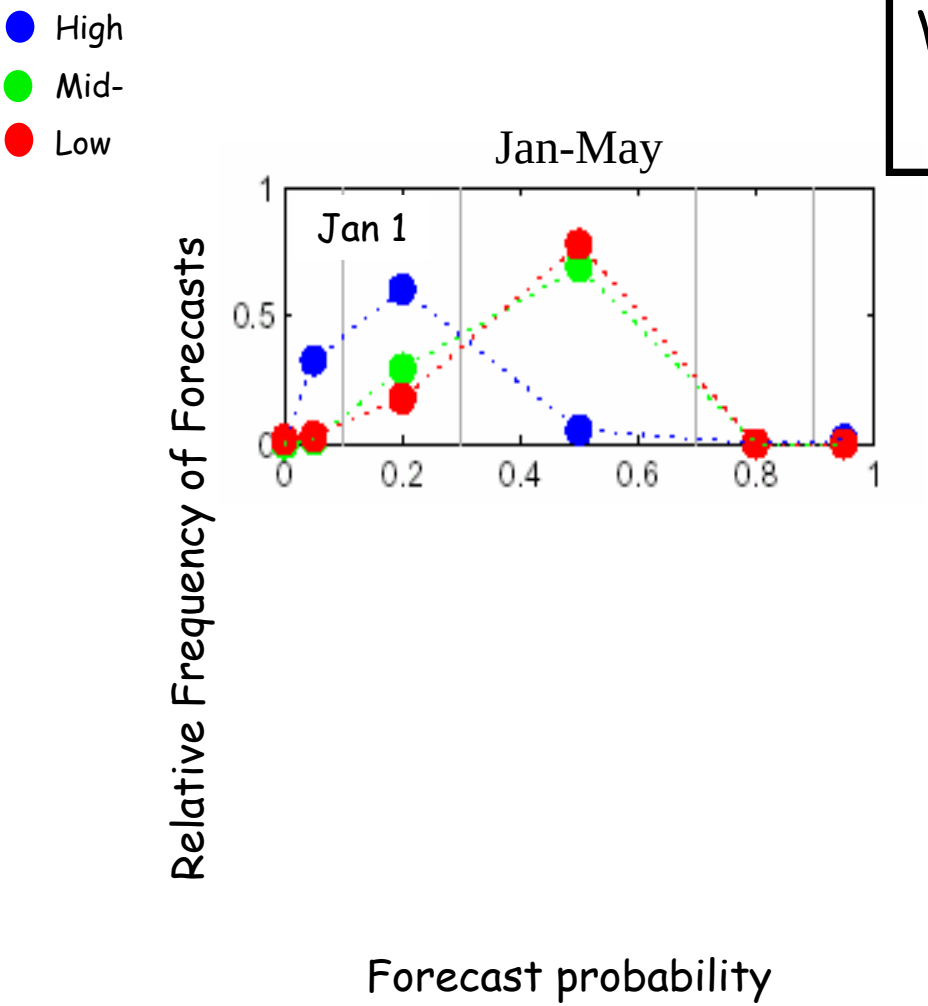
Not much discrimination!



Discrimination: Lower Colorado ESP Supply Outlooks

When unusually low flows happened…

P(F|Low flows)



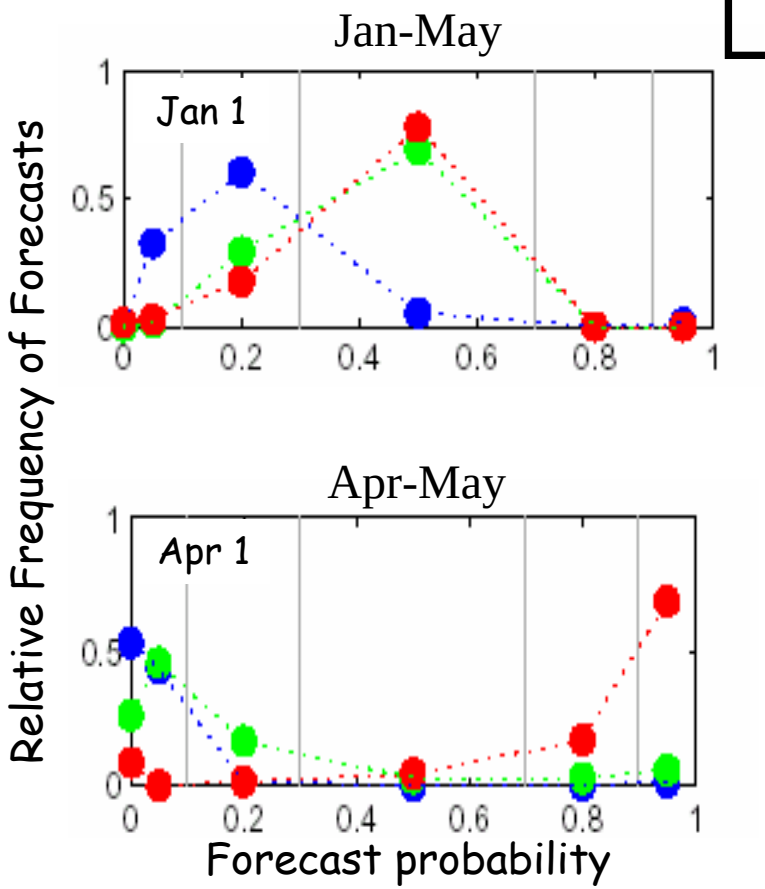
There is some discrimination…

Early forecasts warned "High flows less likely"

From K. Franz (2002)

Discrimination: Lower Colorado ESP Supply Outlooks

High
Mid-
Low



When unusually low flows happened…

P(F|Low flows)

There is some discrimination…

Early forecasts warned "High flows less likely"

Good Discrimination…

Forecasts were saying:

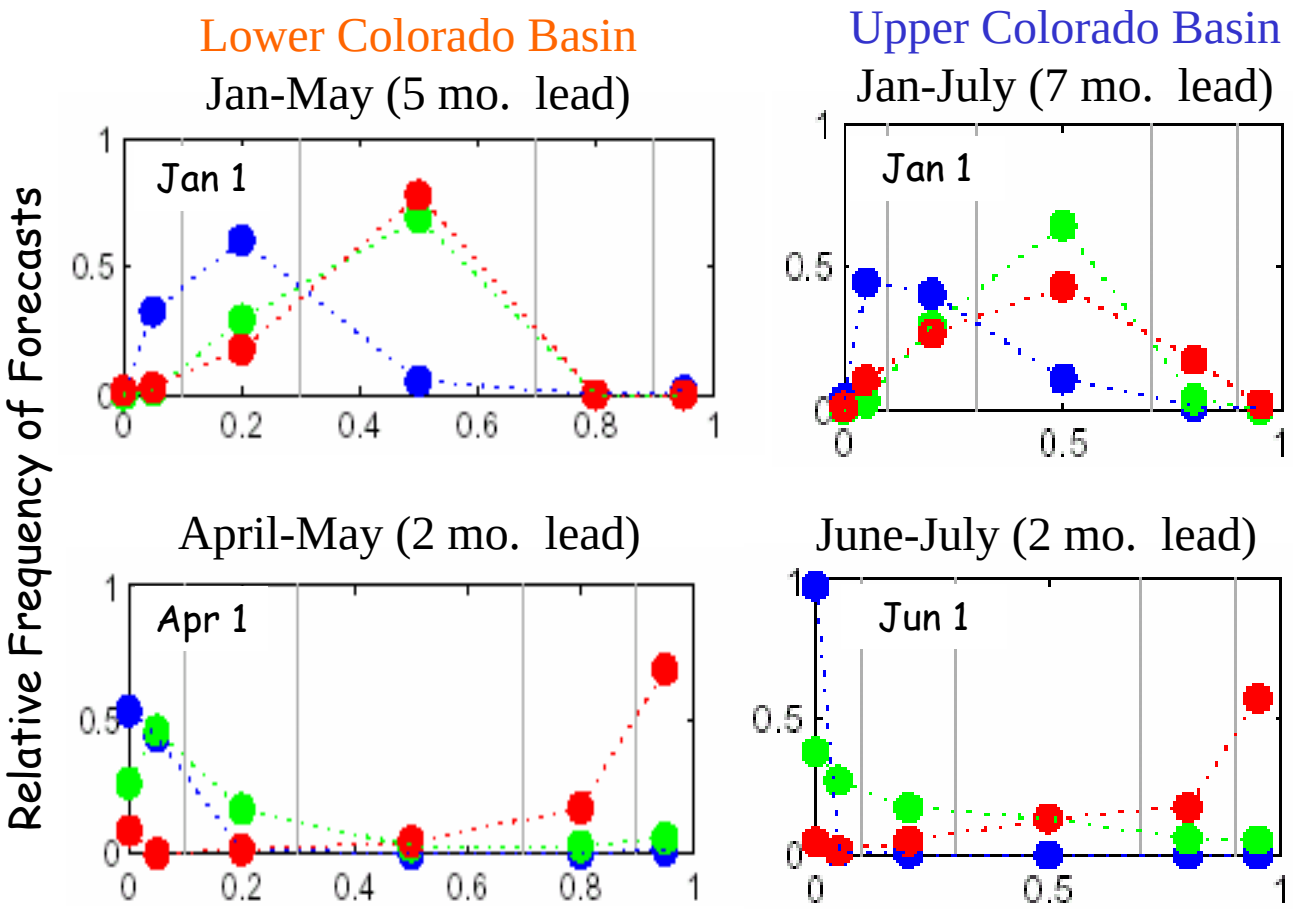
1) high and mid- flows less likely.

2) Low flows more likely

From K. Franz (2002)

Discrimination: Colorado Basin ESP Supply Outlooks

For observed flows in lowest 30% of historic distribution



1)High flows less likely.

2) No discrimination between mid and low flows.

3) Both UC and LC show good discrimination for low flows at 2-month lead time.

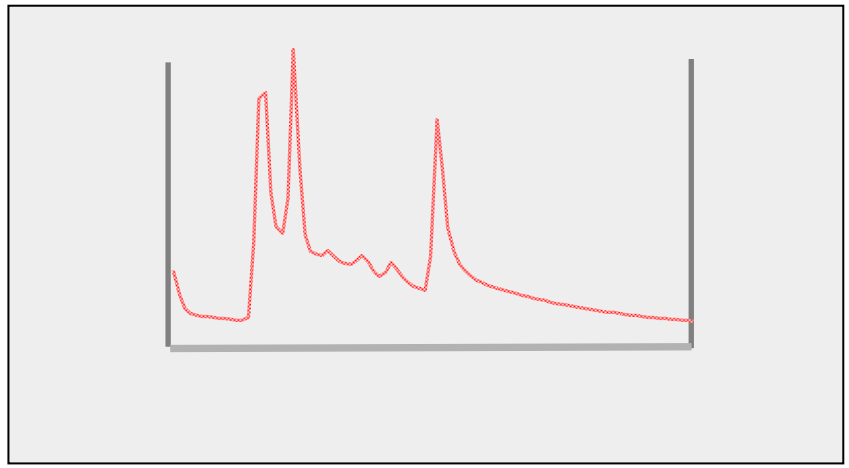
high 30%
mid 40%
low 30%

(Franz, 2001)

Comparing Deterministic & Probabilistic Forecasts

Deterministic forecasts

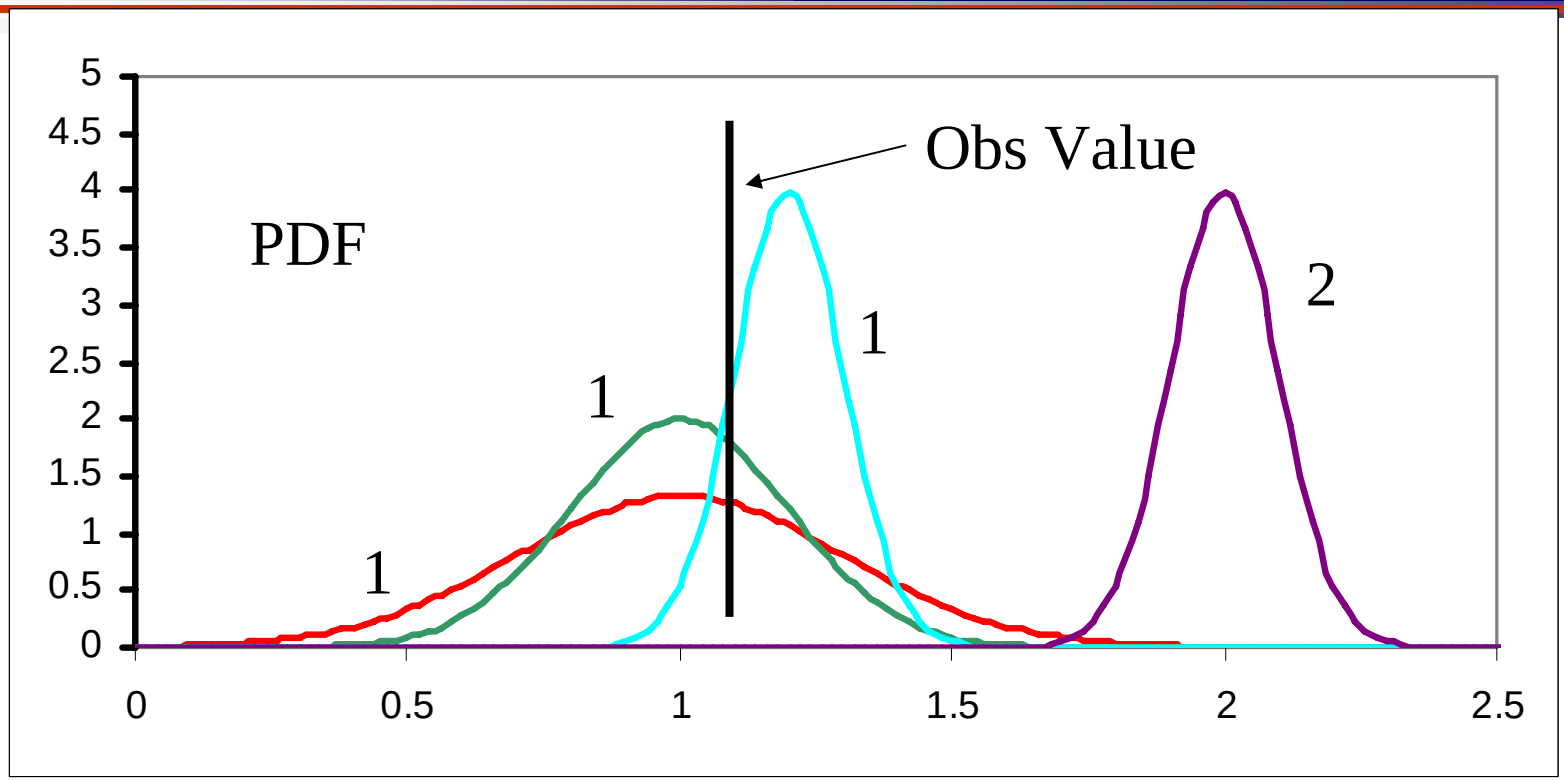
- traditional in hydrology
- sub-optimal for decision making



Common perspective

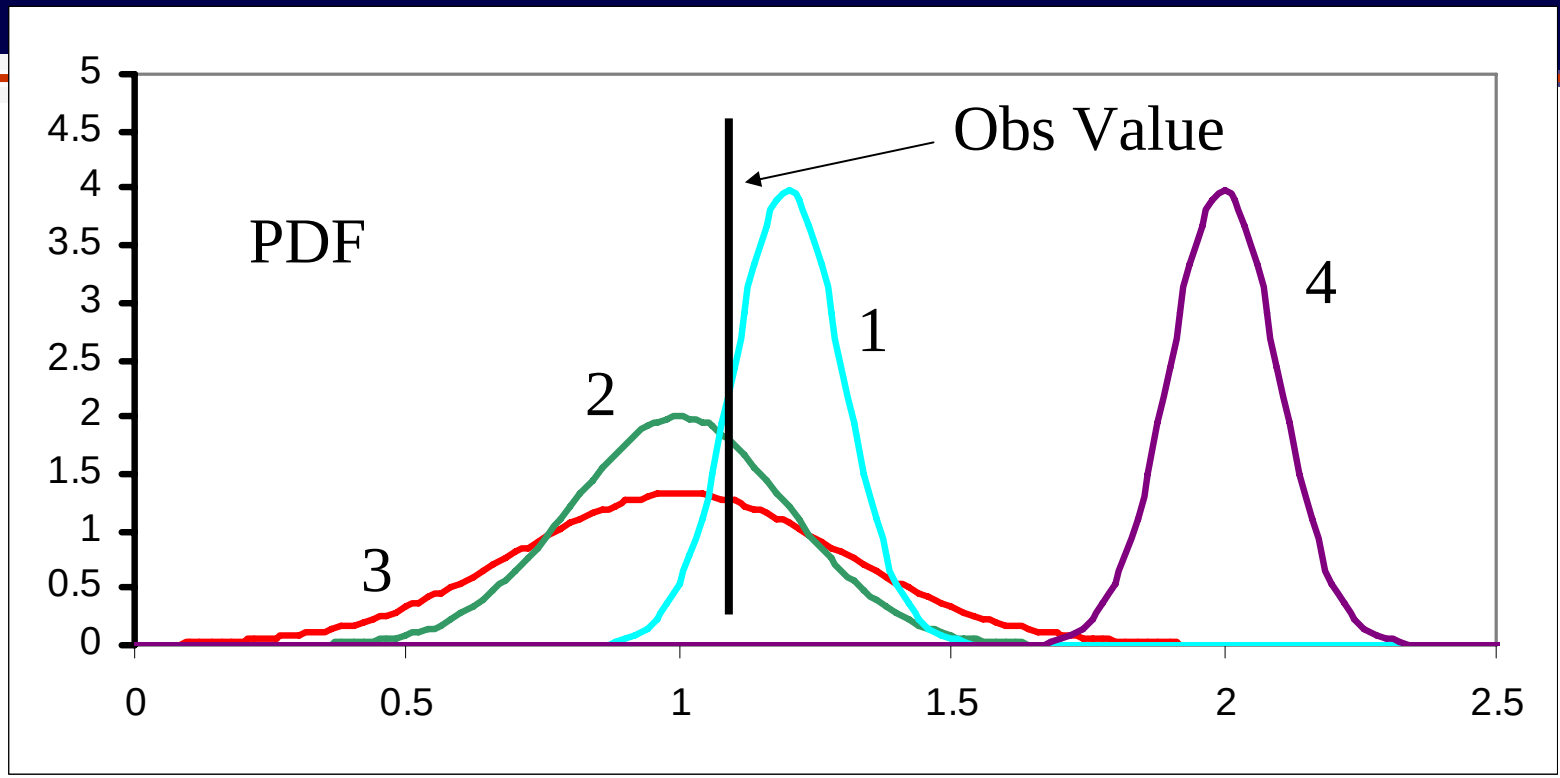
"Deterministic model simulations and probabilistic forecasts … are two entirely different types of products. Direct comparison of probabilistic forecasts with deterministic single valued forecasts is extremely difficult"

What's wrong with using 'deterministic' metrics?



Metrics that use only the central tendency of each forecast pdf will fail to distinguish between red, green, and aqua forecasts, but will identify the purple forecast as inferior.
Example metric: MSE of ensemble mean compared to MSE of long term mean of observations (variance of obs.)

From: A. Hamlet, U. Washington

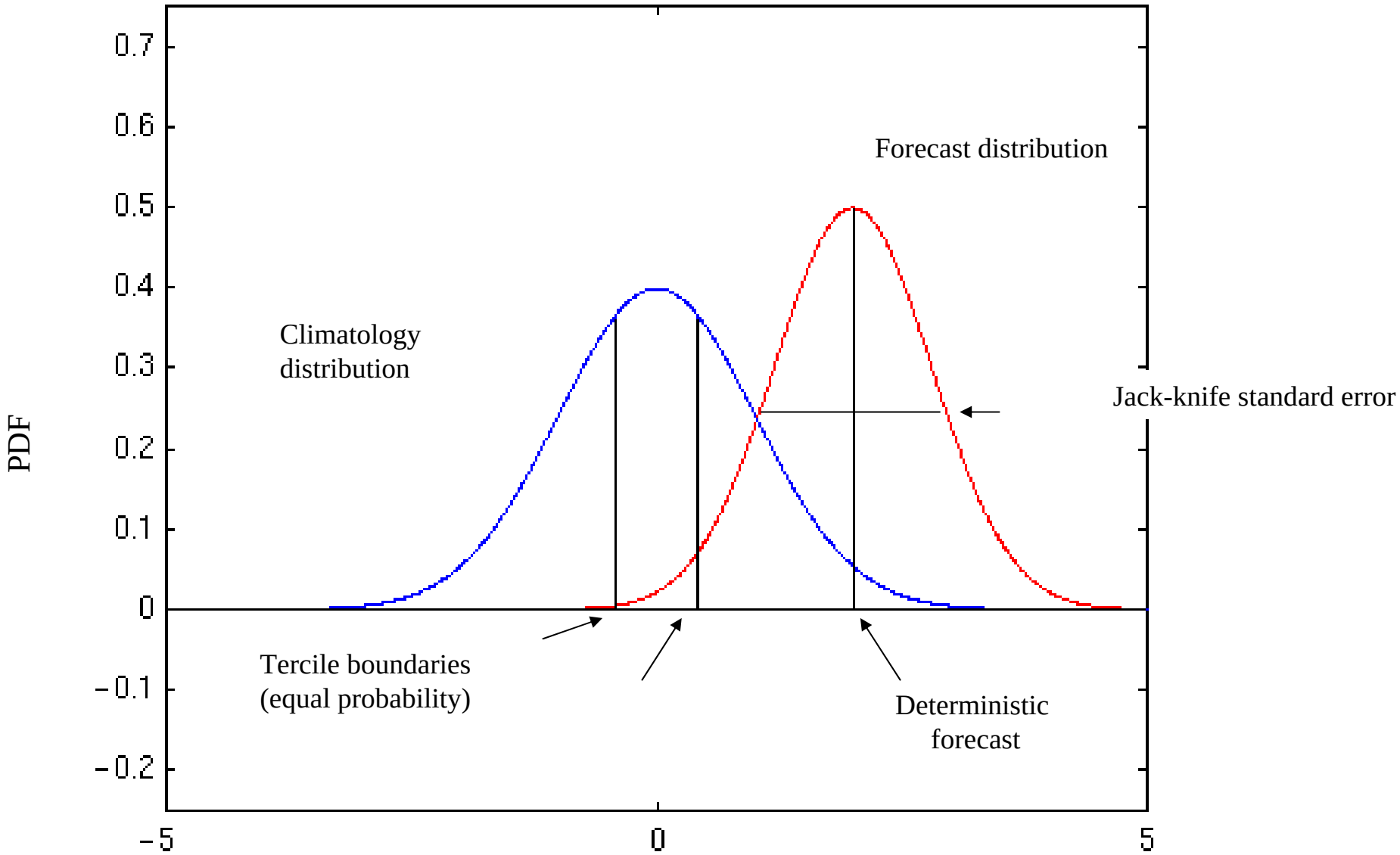


More sophisticated metrics that reward accuracy but punish spread will rank the forecast skill from highest to lowest as aqua, green, red, purple.

Example metric: average RMSE of ALL ensemble members compared to average RMSE of ALL climatological observations.

From: A. Hamlet, U. Washington

Deterministic vs. Probabilistic Forecasts



Forecast Evaluation: Critical Needs

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

Compare by converting deterministic forecasts to probabilistic form.

Better estimation of naturalized flows.

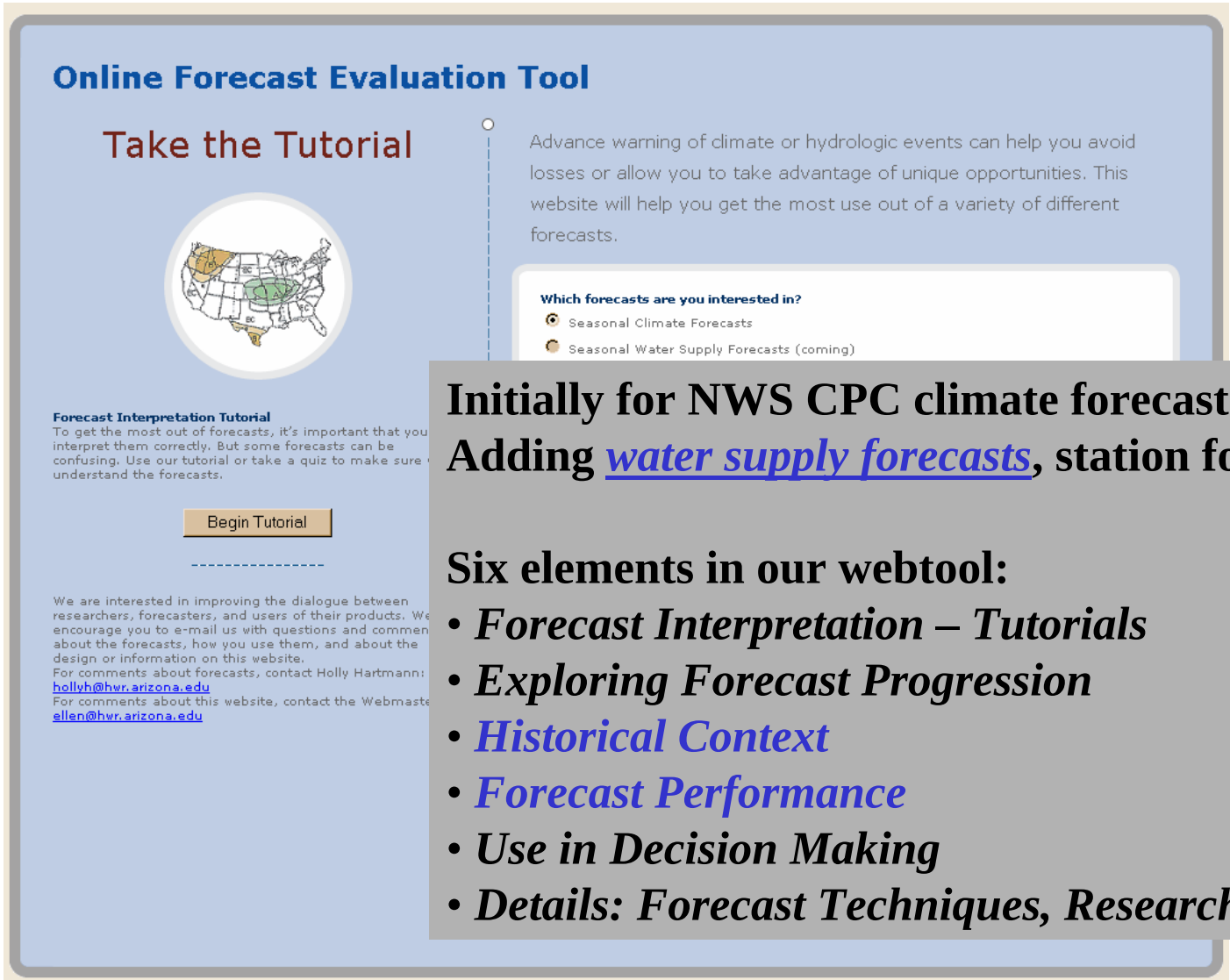
Cooperation of forecasting agencies and groups.

Archives of forecasts and forecasting information.

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

Communication of forecast performance to users.

http://fet.hwr.arizona.edu/ForecastEvaluationTool/



Initially for NWS CPC climate forecasts
Adding *water supply forecasts*, station forecasts

Six elements in our webtool:

- ***Forecast Interpretation – Tutorials***
- ***Exploring Forecast Progression***
- ***Historical Context***
- ***Forecast Performance***
- ***Use in Decision Making***
- ***Details: Forecast Techniques, Research***