Hydrology and Verification

Edwin Welles
DOH/RDM Conference
6/10/04

Outline of This Talk

- History of Hydro Verification
- Purposes for Verification
- Summary of Verification Process with examples
- Local Verification
- National Verification
- Probabilistic Verification

NWS Hydro Verification History

- CR RFC Verification implemented (1982)
- Flash Flood Watch/Warning Verification (1986)
- Dave Morris paper (1988)
- WSOM National Hydrologic Verification Program (1996)
- NRCS report: go do verification (1996)
- National RFC Verification implemented (2001)
- SR RFC Verification implemented (2001)
- General Johnson wants metrics after events (2004)
- New Verification program delivered to RFCs (2004)

NWS Hydro Verification Future

- National RFC Verification includes all locations
- National RFC Verification changes statistics
- River Watch/Warning Verification
- Hydro Verification training course established at NWSTC

Hydro Verification
The Real History

- NWS: Run Away
- Academia: Ignore it

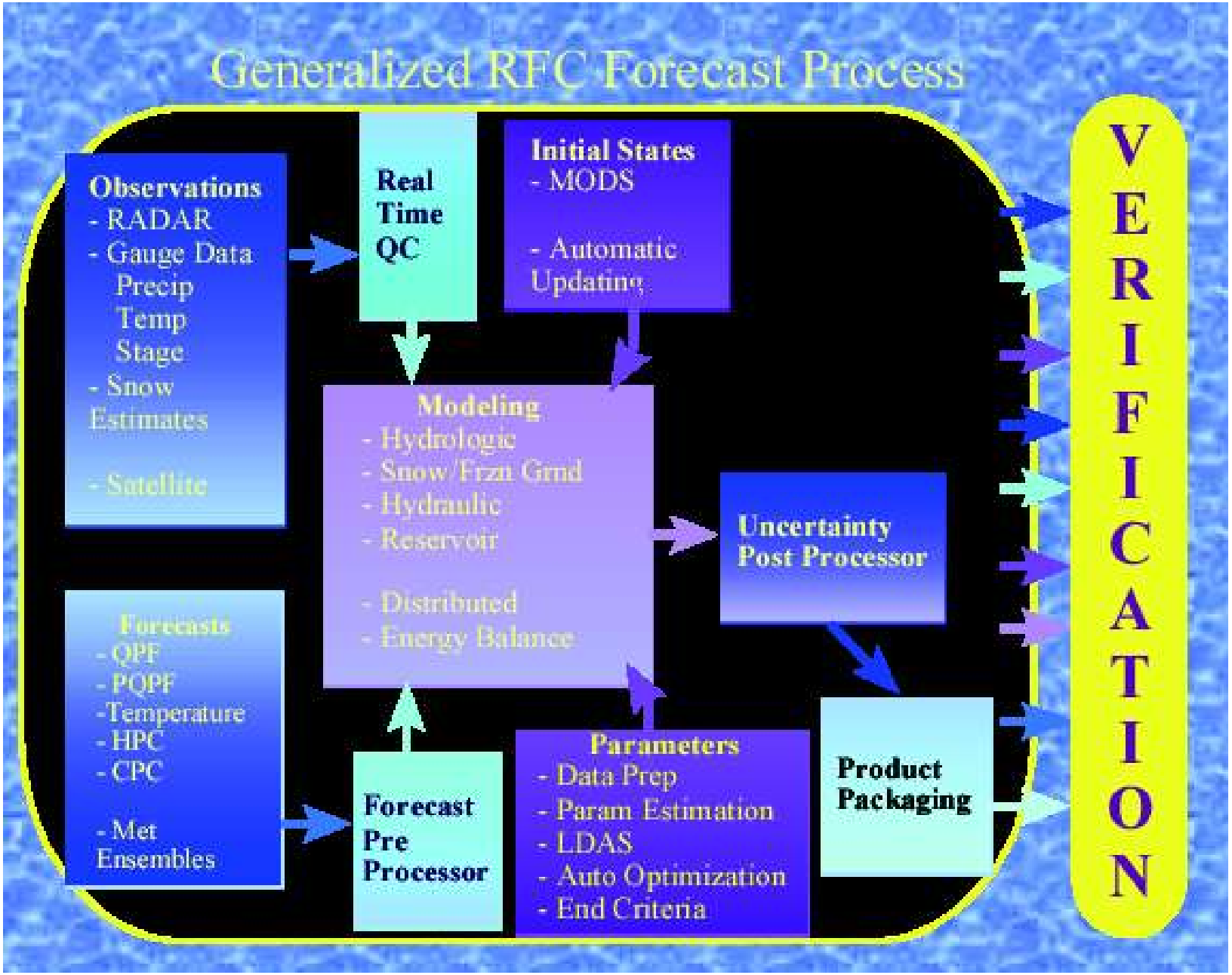
Why Verification is Important

- Verification can help us understand sources of **Skill** in our forecasts.
- Verification can help us understand sources of **Uncertainty** in our forecasts.
- Verification can help us understand the **Conditions** when we are or are not skillful.
- Verification can help us demonstrate the value of our work by **Tracking** changes in skill.

The Purposes of Verification
Scientific or Local Verification

- Studying the forecasts to determine when and why they are skillful and not skillful
  - Identifying limits of predictability
  - Determining how to make the forecasts better
- Defining how forecast process updates will improve forecasts
- Akin to Calibration
  - Except we are calibrating the entire forecast system

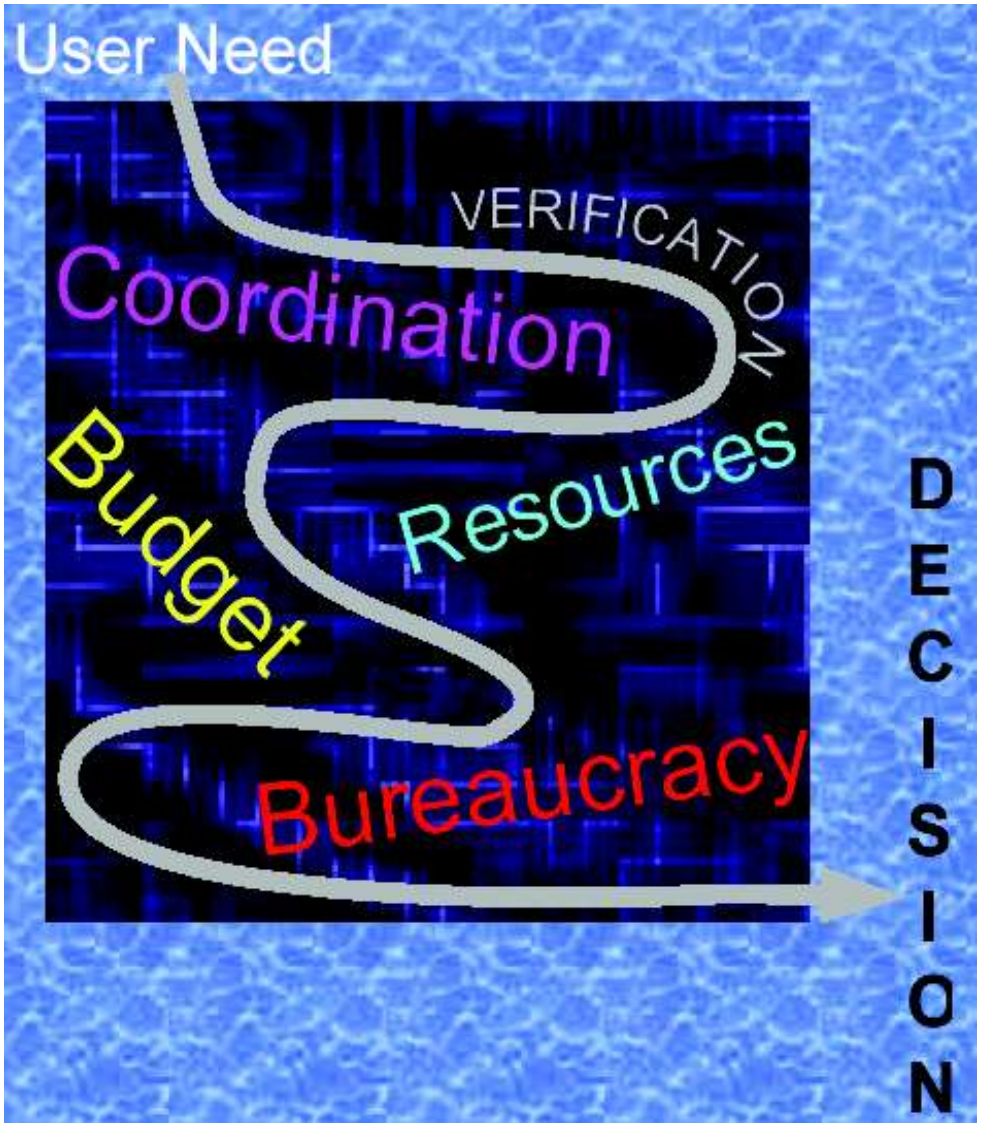
Verification Closes the Loop



Purposes of Verification cont...
Administrative or National Verification

- General health of the entire service
  - Includes timeliness and number of forecast locations as well as forecast accuracy
- Used to communicate value to non-scientists
  - Especially those who fund us.
- Must be supported by effective Scientific verification to be meaningful.

Verification Helps with Decisions



So you want to verify your forecasts

- You need to choose what you will verify,
- You need pairs of forecasts and observations.
- You need a control forecast (or two).
- You need to sort your pairs into informative subsets.
- And you need to select some statistics to characterize those subsets.

What will you verify

- Choose a set of forecasts
  - It needs to be large enough so the metrics are correctly estimated, but the sample must be homogeneous.
    - Same predictand
    - Similar lead times
    - Similar time horizons
    - Similar forecast processes
- Pooling to many forecast types will make it difficult to interpret the metrics you compute.
  - Make it more likely changes in the metrics are not related to changes in forecast skill.
  - Bias results to most commonly sampled regime

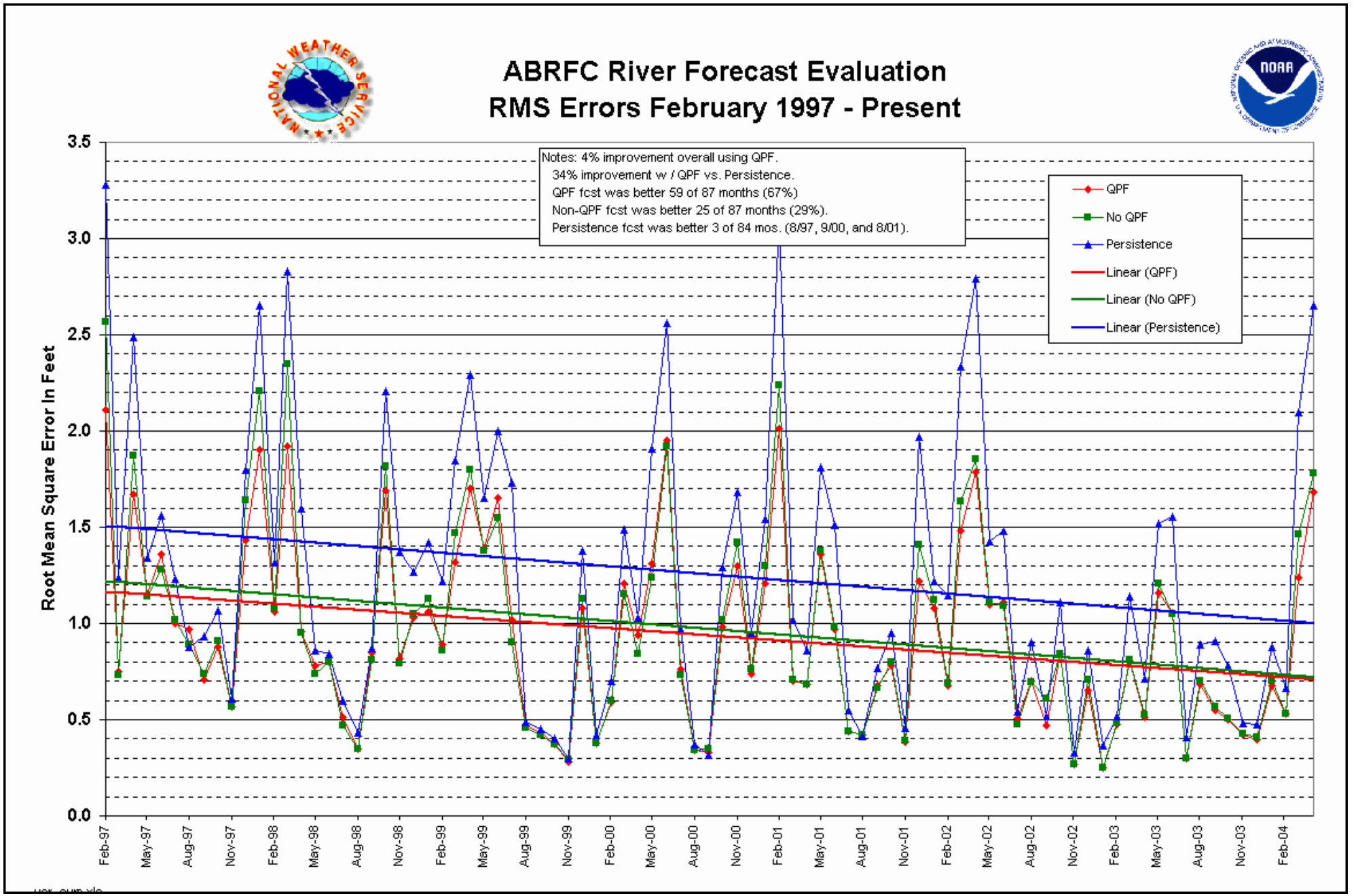
Pairs of forecasts and observations

- Match them within some window
- The Big Miss: a flood with no forecast.
  - Flood Only locations
- Must count these missed forecasts
  - Difficult for mean error type metrics
  - Can be done for categorical metrics

Why We Need Control Forecasts

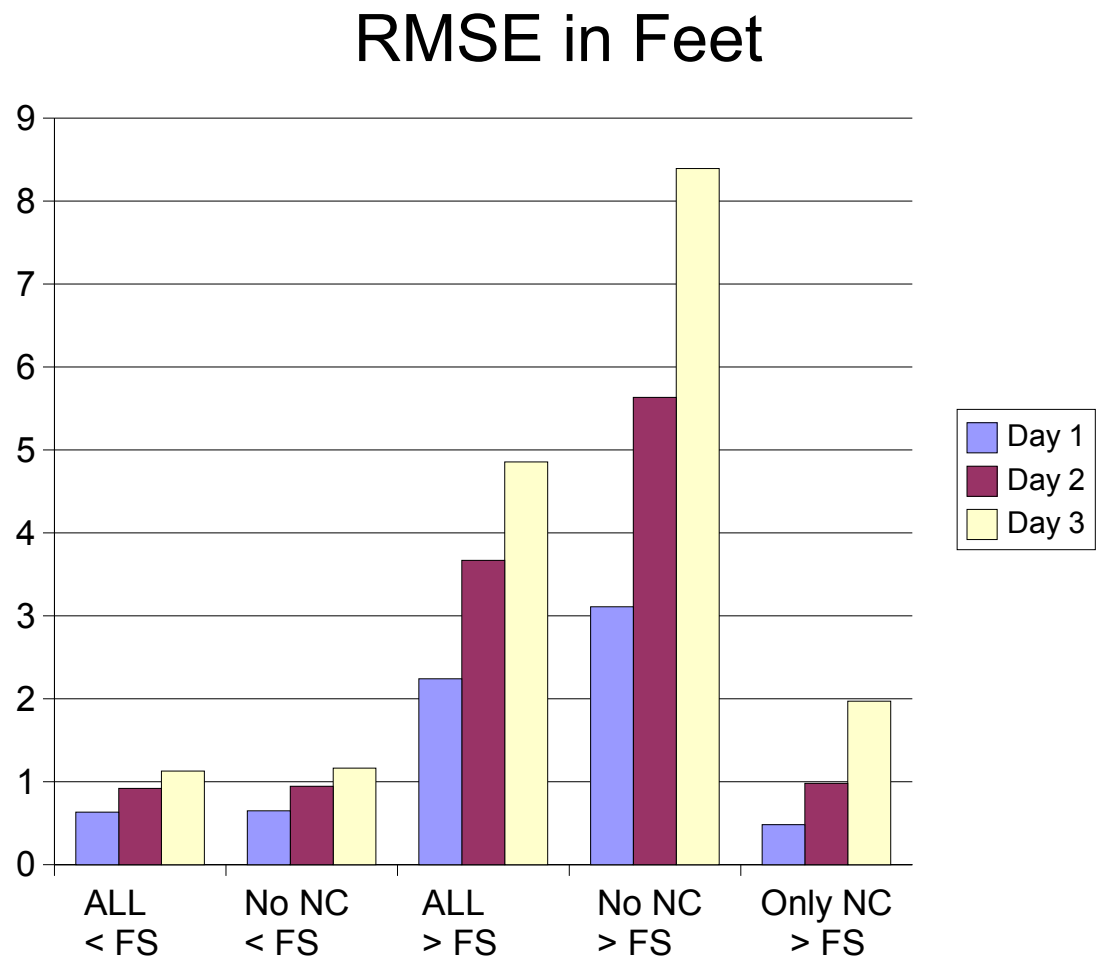
- They provide us a perspective to help us understand the vagaries of the forecast skill
- Help us relate different locations and different periods.
- Help us identify sources of skill and error.
- 3 examples follow.

The Value of a Control Forecast

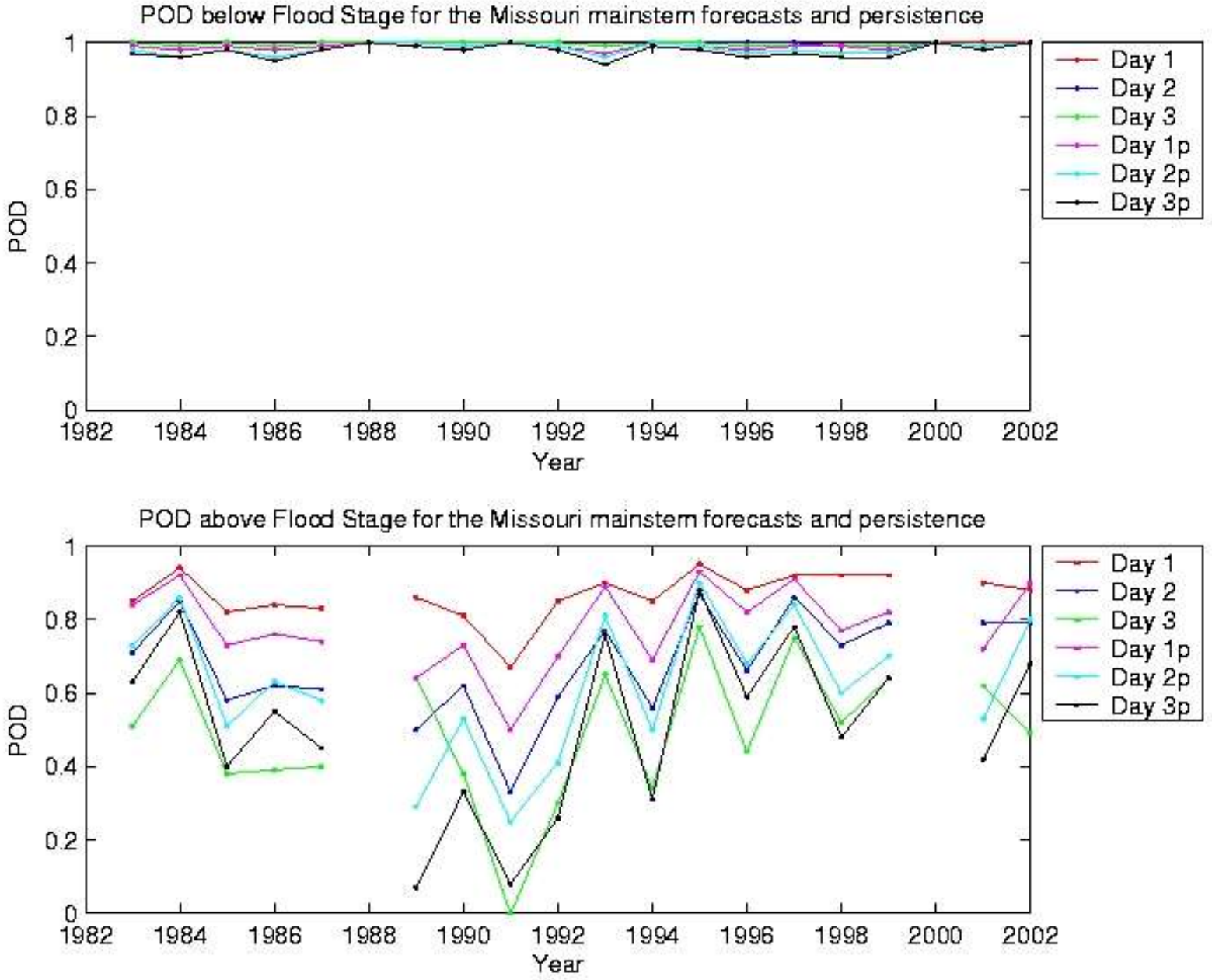


National Statistics 4/2001 to 10/2003 Fast Response Locations

- NCRFC contributes 60% of the Samples
- Very low RMSE for two NC locations
  - 1 mis-classified
  - 1 ???????
- Need a control forecast to analyze further.

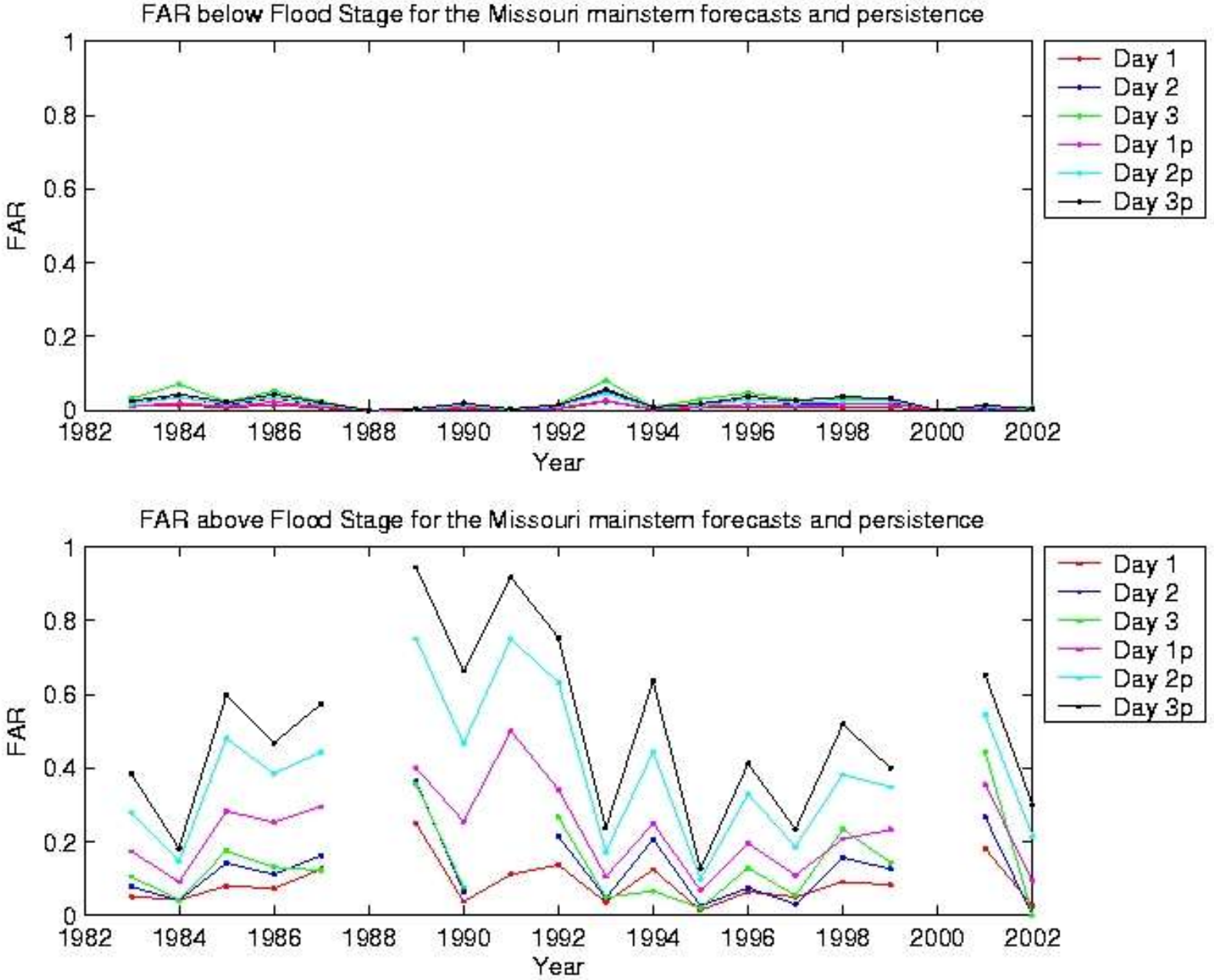


POD for Missouri Mainstem Locations



Figure(18): Annual POD for Missouri mainstem forecasts and persistence, days 1, 2, 3.

FAR for Missouri Mainstem Locations



Figure(28): Annual FAR for Missouri mainstem forecasts and persistence, days 1, 2, 3.

Informative Subsets

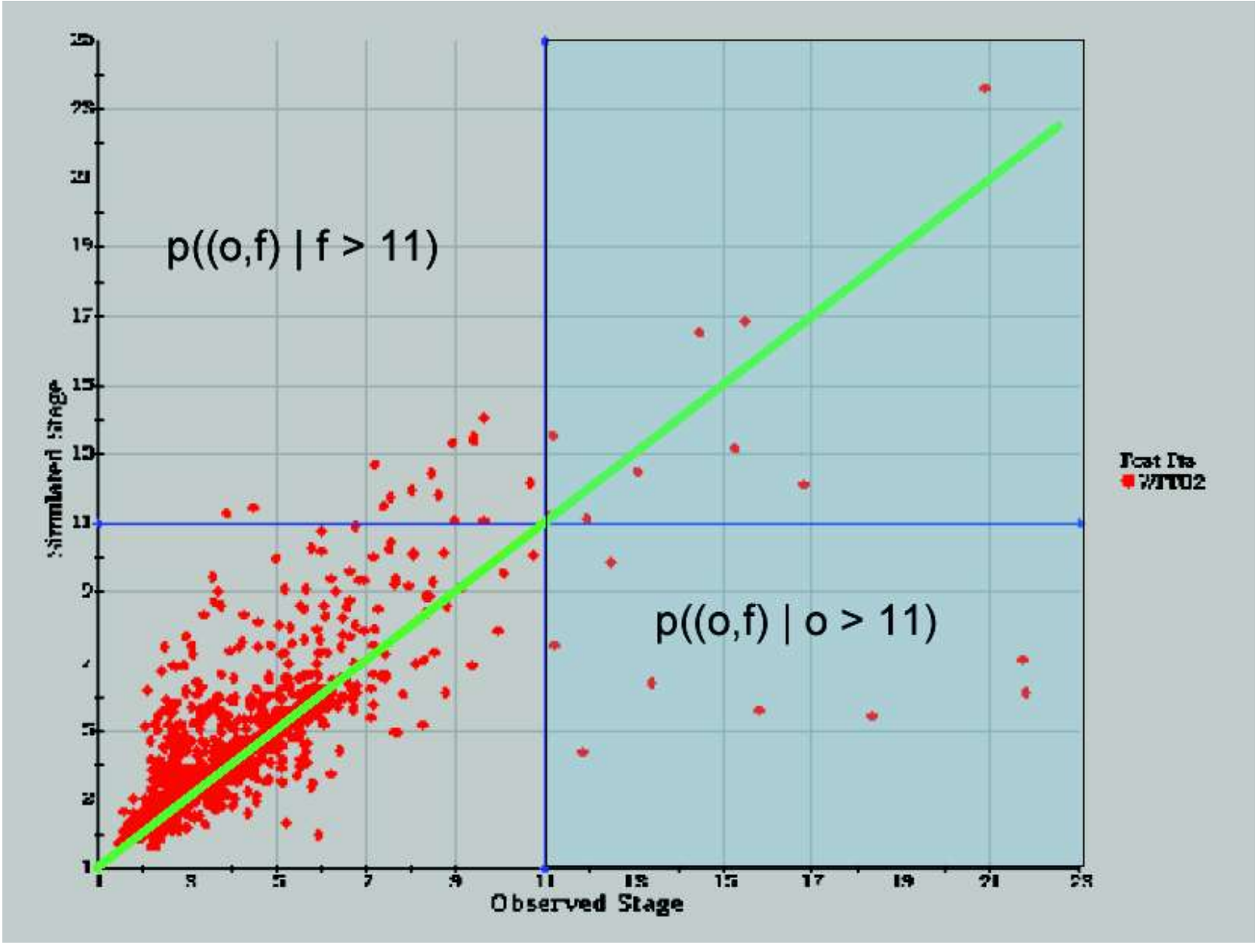
- The sorting depends upon your goal.
  - More sorting for scientific (local) verification than administrative (national)
- There is a whole theory about subsets
  - Distributions oriented verification
- Other things to sort by
  - Lead-time
  - Gradient
  - Season
  - Basin type
  - Forecast Situation (e.g. backwater)

Two Characteristics of Forecasts

- Discrimination
  - Do the forecasts discriminate between types of future events?
  - If a flood happened was there a forecast?
  - Sort based upon the observed values.
- Reliability
  - When we forecast an event, are the forecasts reliable?
  - If we forecast something, does it happen?
  - Sort based upon the forecast values.

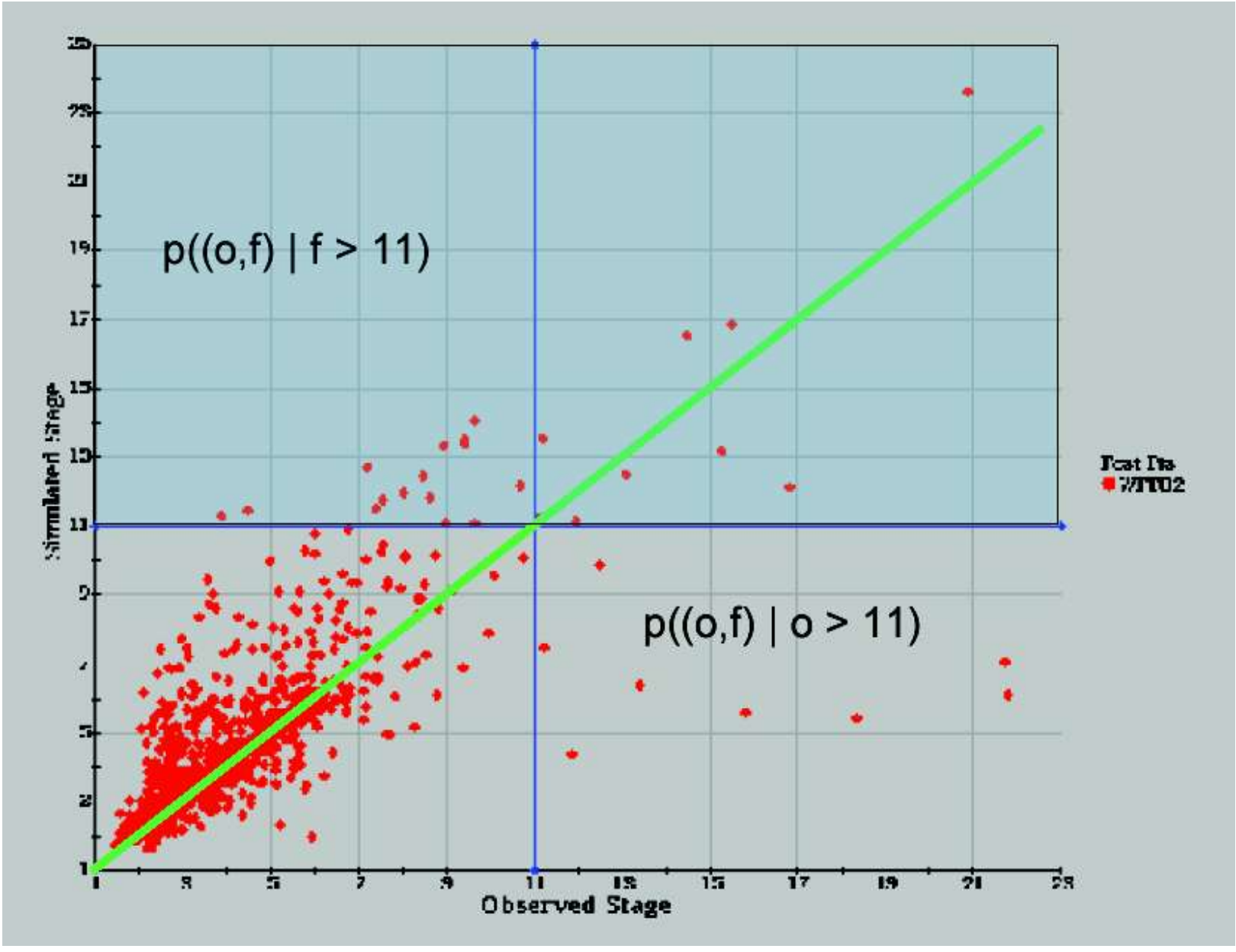
p((o,f) | o) vs. p((o,f) | f)

Discrimination



p((o,f) | o) vs. p((o,f) | f)

Reliability



Selecting some statistics

- Categorical metrics
  - POD, FAR, CSI, GS
- Accuracy measures
  - RMSE, ME, MAE
- Non-dimensional scores
  - Skill scores, Correlation
- Probability measures
  - RPS, RPSS, PS

Contingency Table for Categorical Forecasts

Forecasts \ Observations	Yes	No
Yes	A	B
No	C	D

Note: You can convert continuous and probability forecasts to categorical with a threshold.

- Probability of Detection
  POD=A/(A+C)
- False Alarm Rate
  FAR=B/(A+B)
- Critical Success Index
  CSI=A/(A+B+C)
- Probability of False Detection
  POFD=D/(B+D)
- Peirce Skill Score
  PSS=(ad-bc)/(a+c)(b+d)
- Gerrity Skill Score
  Average PSS

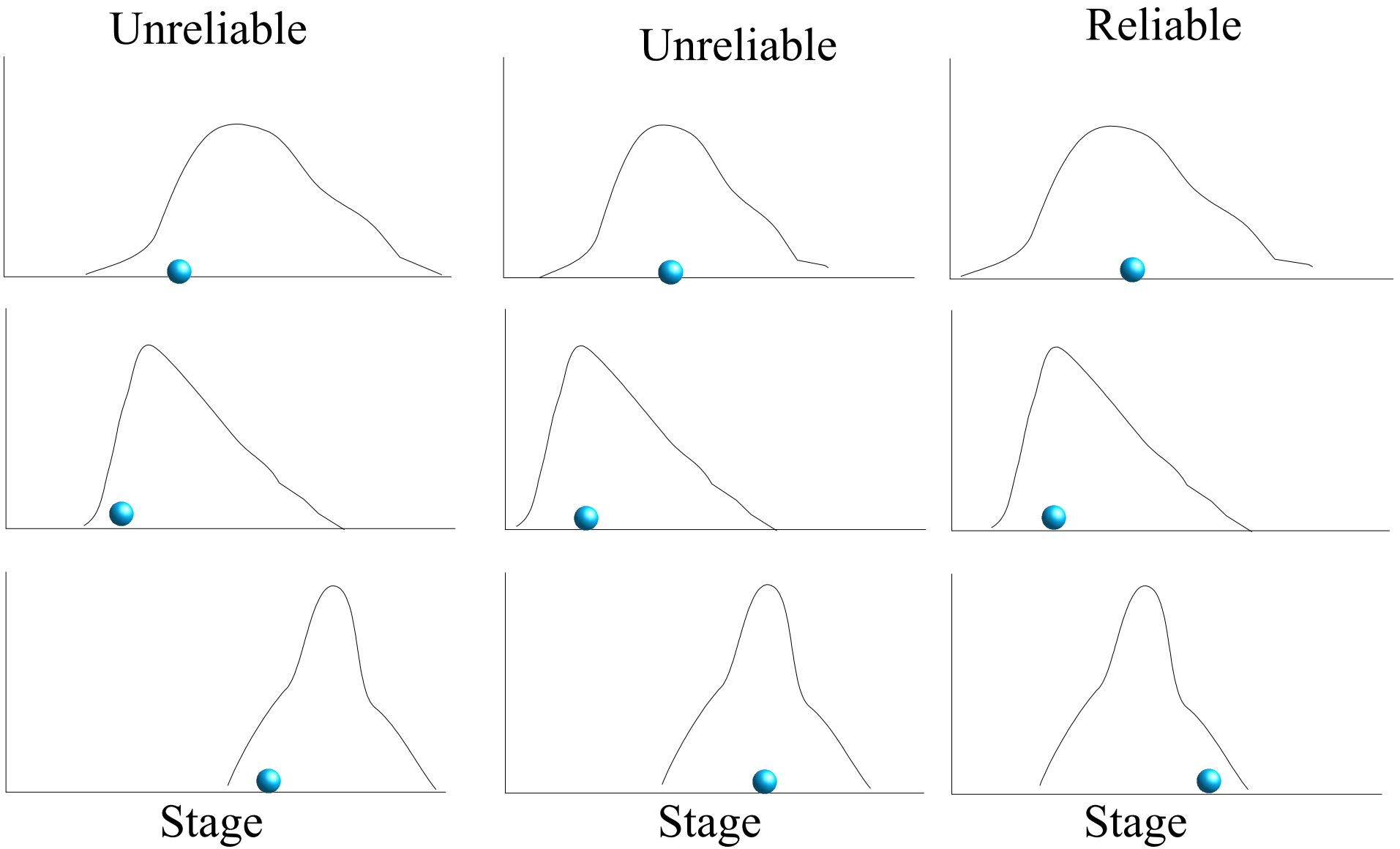
Non-Dimensional Measures

- Skill Score: measures performance relative to some control.
  - $$SS = \frac{(S_{cntl} - S_{actual})}{(S_{cntl} - S_{perfect})}$$
- Correlation Coefficient: measures linear correlation between variables
  - $$R = \frac{cov(obs, fcst)}{\sqrt{var(obs) * var(fcst}}$$

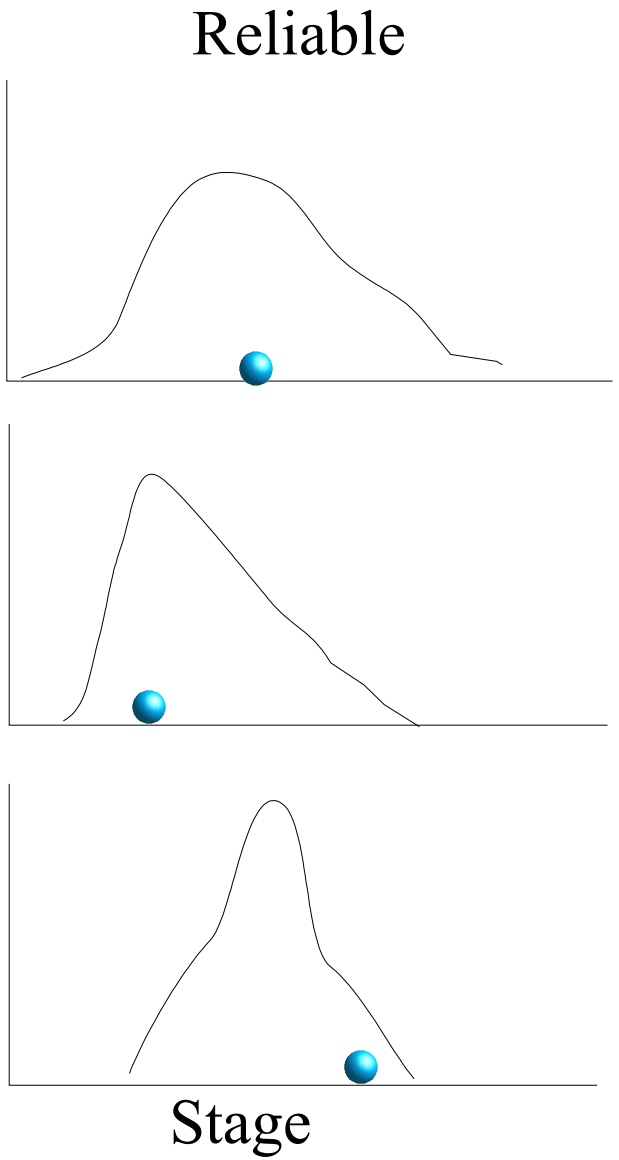
Characteristics of Probability Forecasts
Resolution and Reliability
and Sharpness

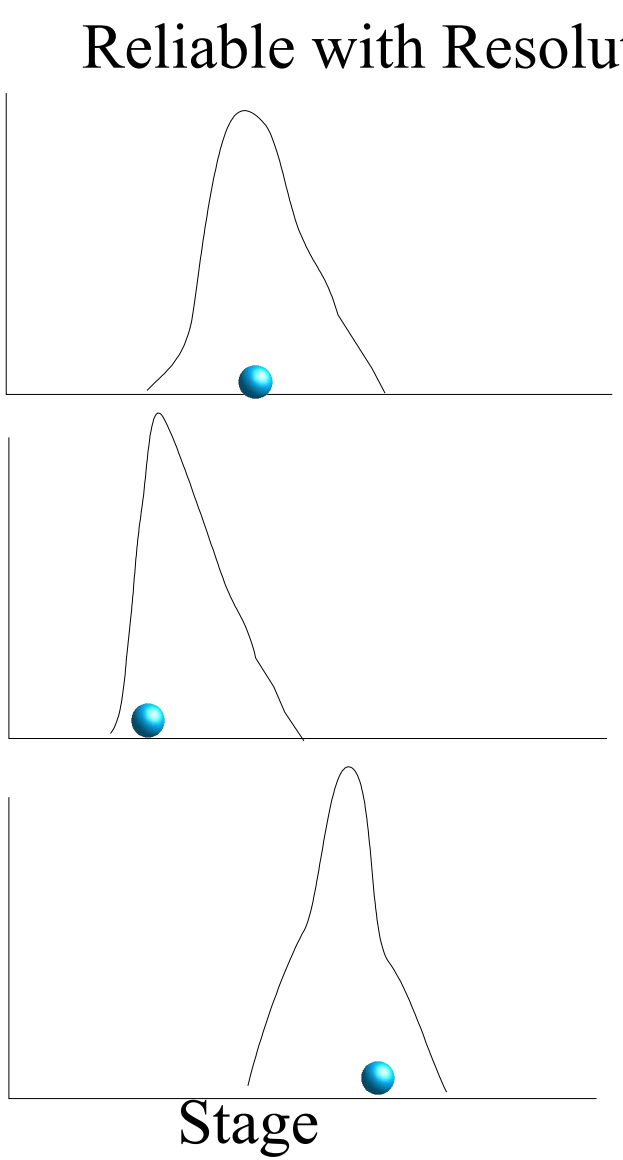
- Resolution
  - Do the forecasts distinguish between upcoming events?
- Reliability
  - Do the forecast probabilities correctly reflect the future uncertainty?
- Sharpness
  - Do the forecast probabilities cluster around 0 and 1, not the mean?

A Reliability Example



And Now with Resolution

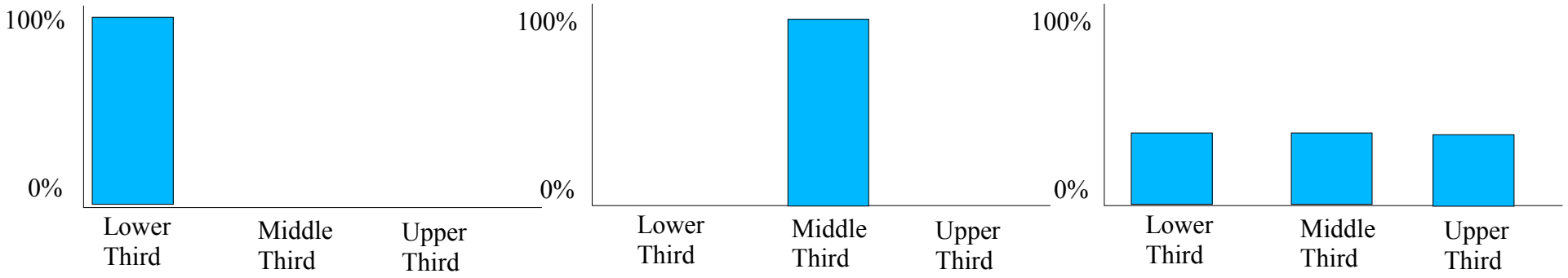


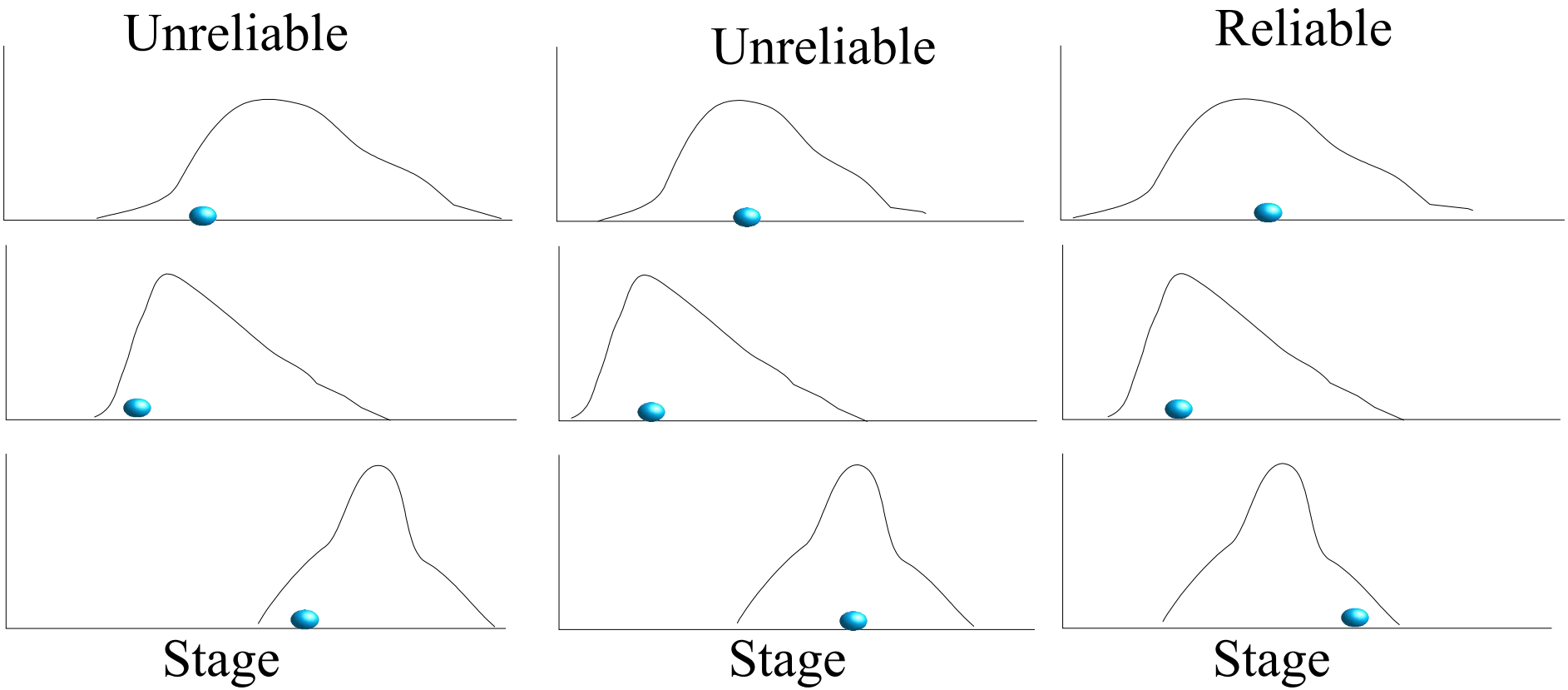


Measuring Reliability

- Rank Histograms
  - Evaluate the frequency the observed falls into probability intervals.
  - We usually compute a Rank Histogram using the Cumulative Distribution Function
  - Call it a “Cumulative Rank Histogram”.
- Reliability Diagrams
  - Evaluate the frequency the observed falls into probability intervals given a forecast for a specific event.

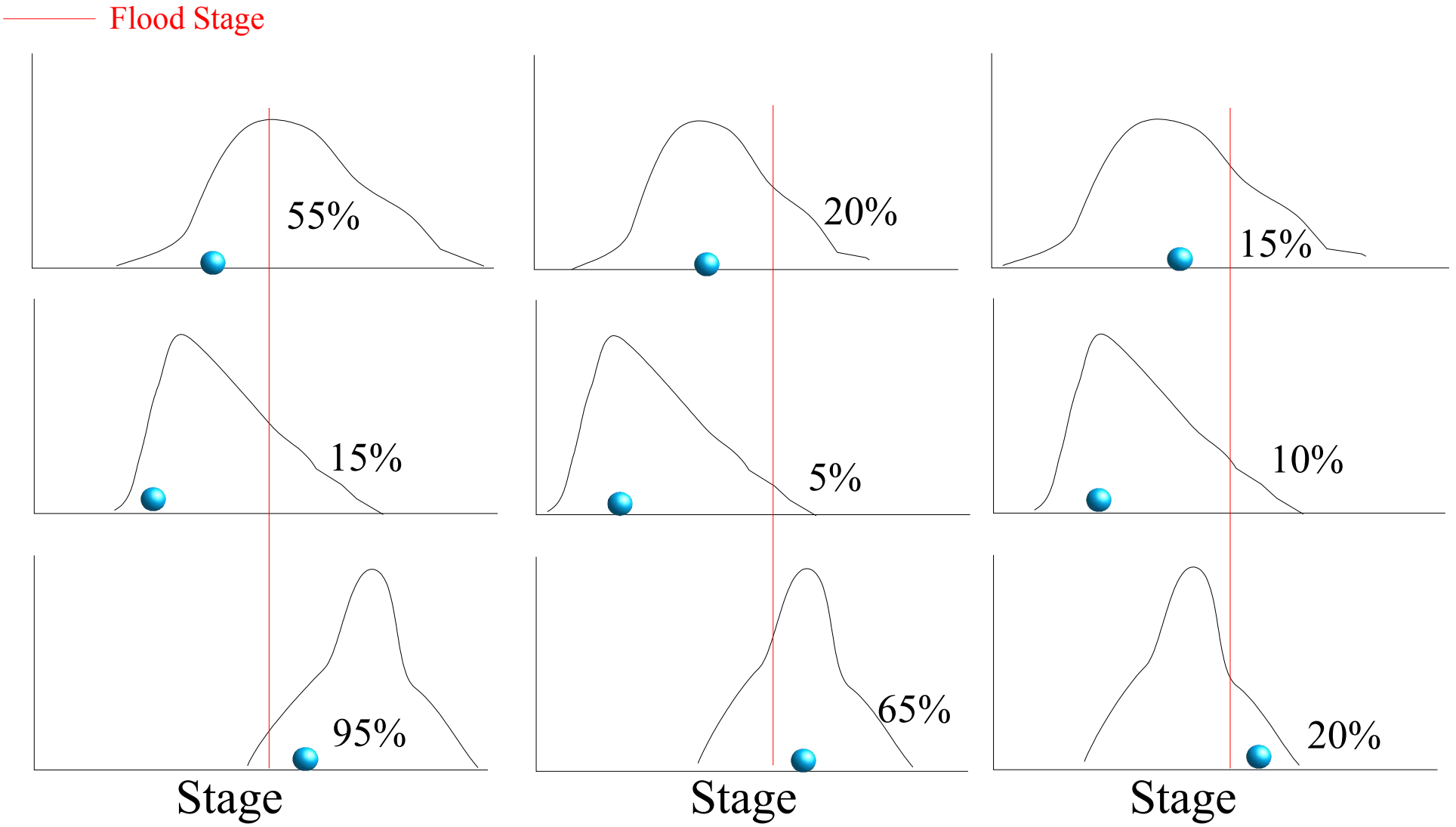
Rank Histograms





We Need to Look at Forecasts of Something

Forecasts of Flooding



Reliability Diagrams

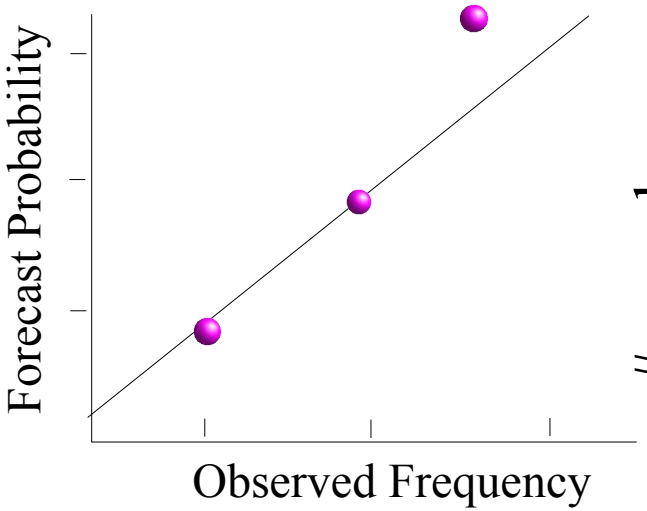
Forecasts of Flooding

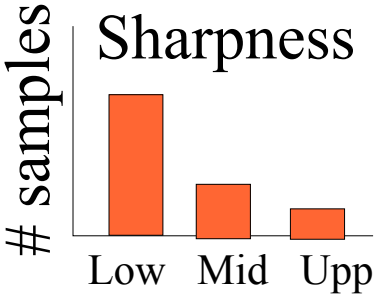
Forecast chance of a Flood

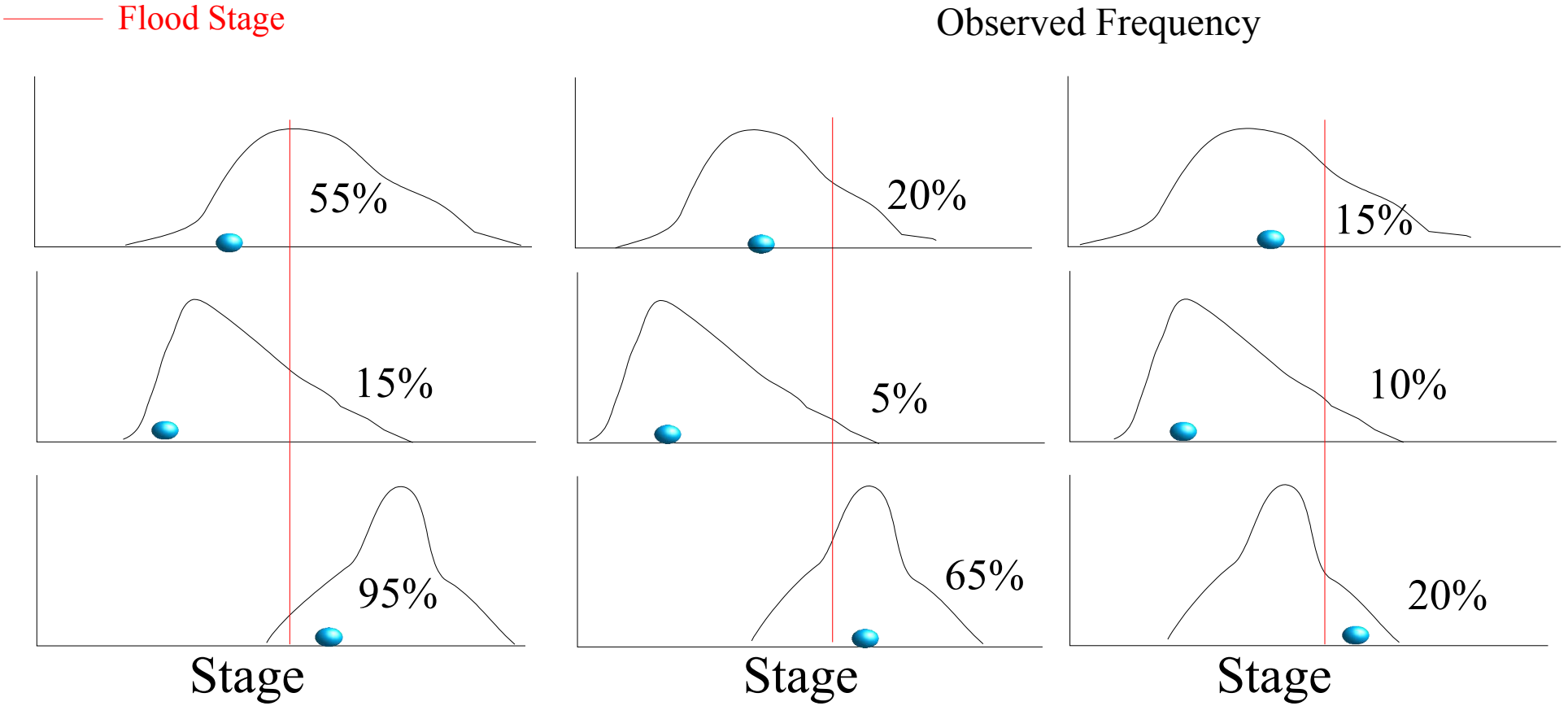
< 33 % : 6 fcsts, 1 obs

33% to 66 %: 2 fcsts, 1 obs

>66%: 1 fcst, 1 obs







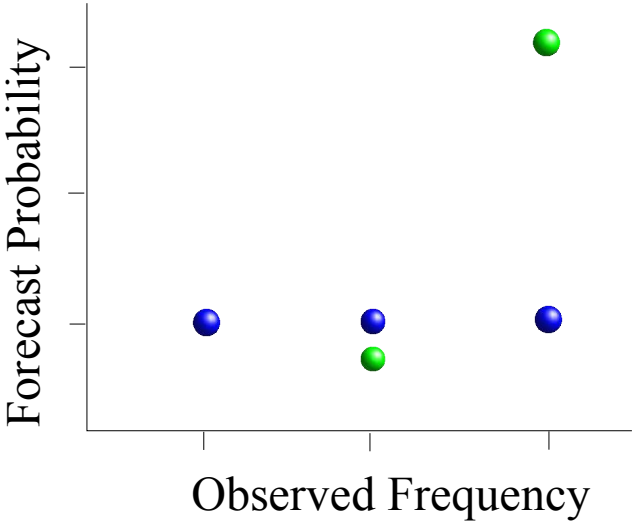
Measuring Resolution

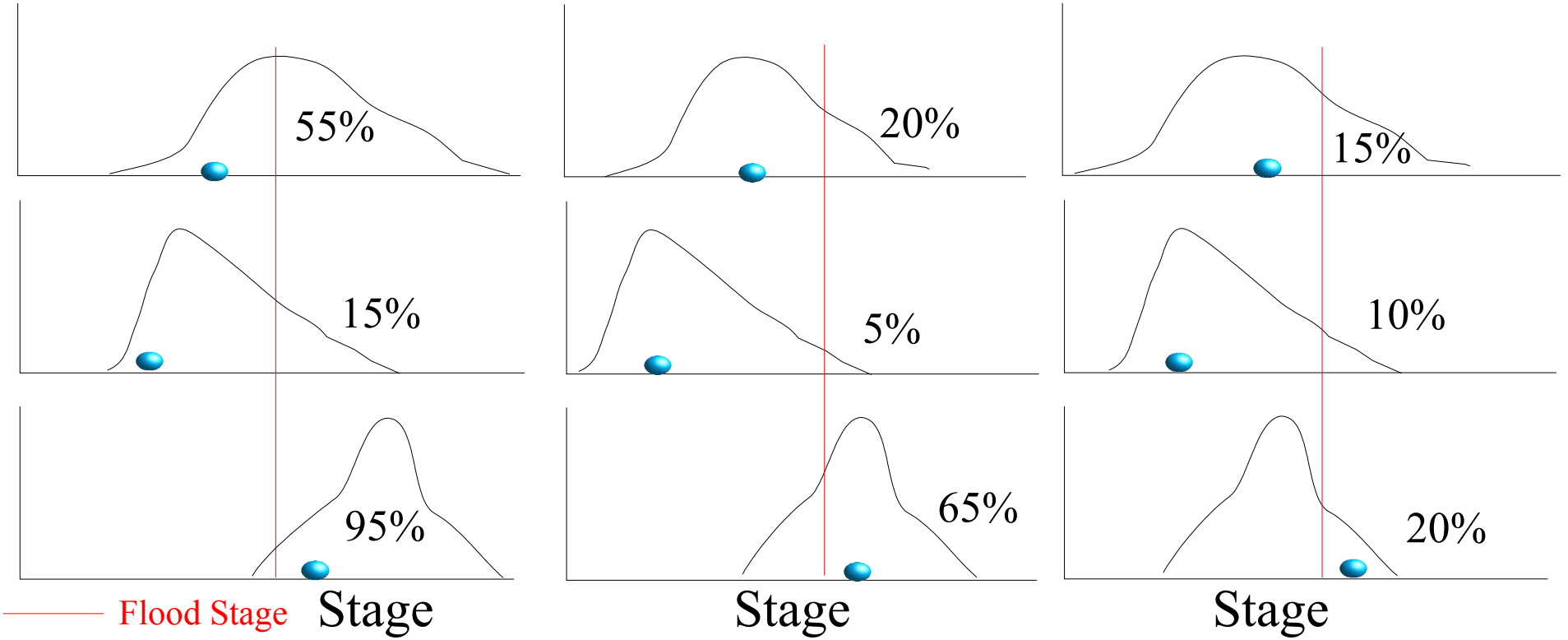
- Discrimination Diagrams
  - Like Reliability Diagrams, but sorting by the observation
  - Look at what the forecast was, before a flood.
- Relative Operating Characteristics (ROC) Diagrams
  - Connects to theory to assess value and to deterministic scores.
- Ranked Probability Score and Brier Score
  - Consist mostly of Resolution

Discrimination Diagrams

Flood and No Flood Events

	Forecast Prob of Flood		
	High	Med	Low
3 Floods	1	1	1
	Forecast Prob of No Flood		
	High	Med	Low
6 No Flood	5	1	0





Relative Operating Characteristics (ROC) Diagrams for Flood Forecasts

>15% Flood

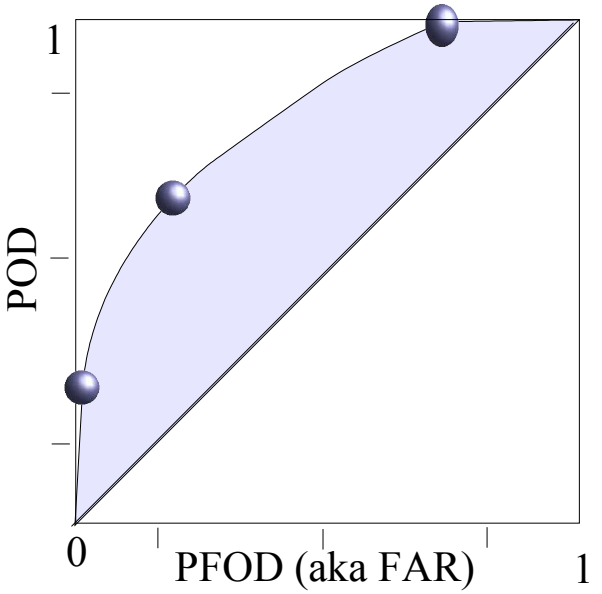
FCST \ OBS	F	NF
F	3	4
NF	0	2

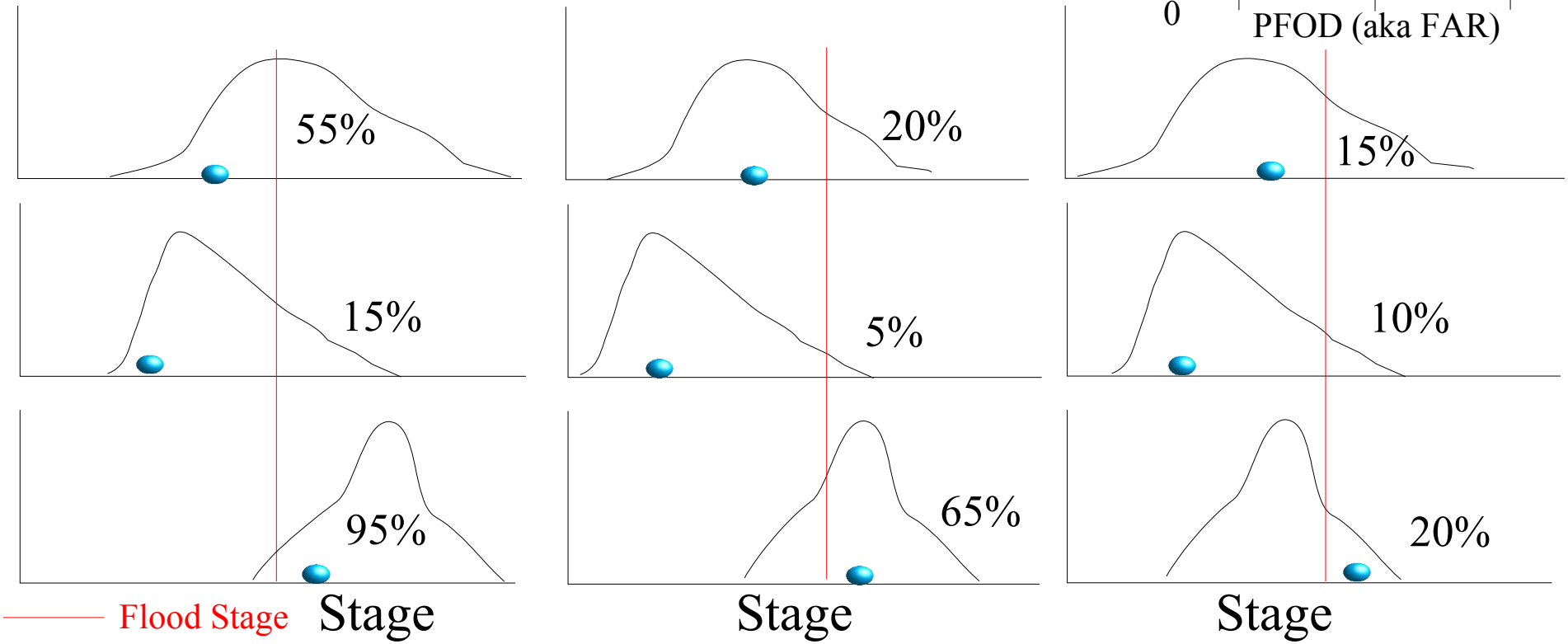
>50% Flood

FCST \ OBS	F	NF
F	2	1
NF	1	5

>85% Flood

FCST \ OBS	F	NF
F	1	0
NF	2	6

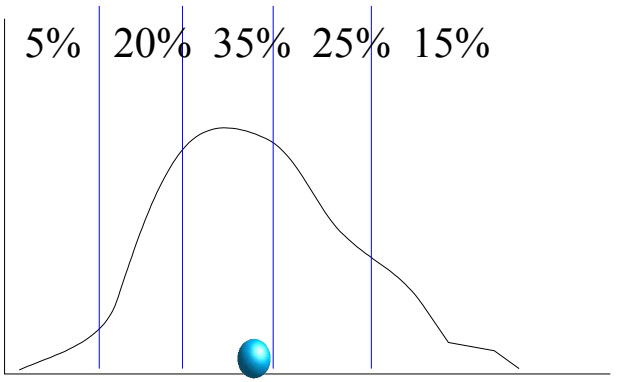




Ranked Probability Score (RPS) for Flood Forecasts

$RPS = 1/k(f_k - o_k)^2$

Where f=0 if the event did not occur
and 5=1 if the event did occur
and k is the category number



K=5

$$RPS = ((0 - 0.05)^2 + (0 - 0.25)^2 + (1 - 0.60)^2 + (1 - 0.85)^2 + (1 - 1.00)^2)/5$$
$$= 0.05$$

For multiple forecasts, take the average RPS.
If K=2; RPS = The Brier Score

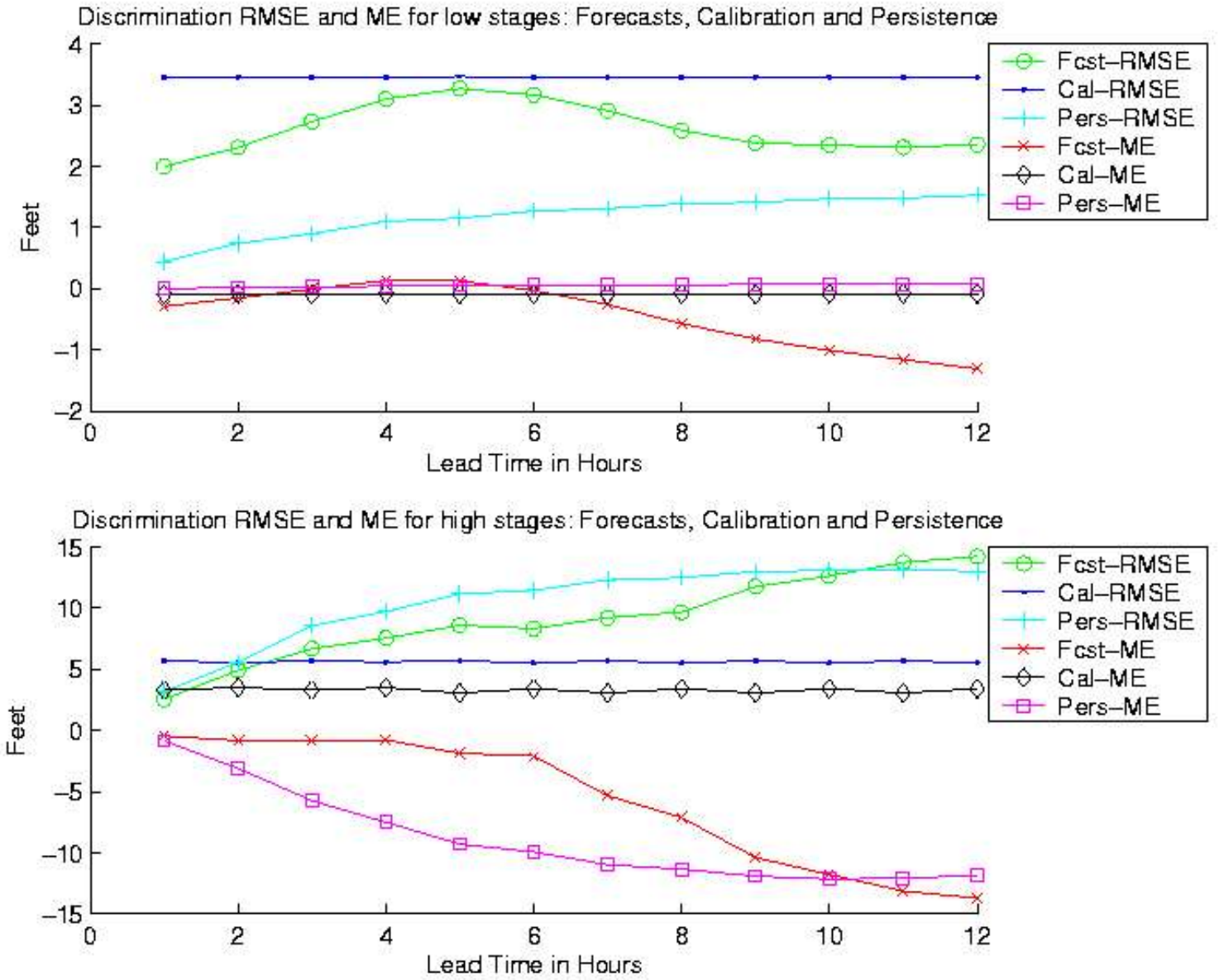
Local RVF Verification: Current and end state

- All issued forecasts stored in your archive db.
- All input forecasts stored in your archive db.
- Two control forecasts (future)
  - Persistence
  - No-mods, observed precipitation and temperature etc
    - Reservoir forecasts (?)
- Numerous sorting options
  - Lead-time, Season, Basin, Forecast type (e.g. with, without QPF), Discrimination, Reliability

Local RVF Verification: Current and end state, cont ...

- Statistics
  - POD, FAR, CSI, RMSE, MAE, ME, Sample Size
  - RMSE skill score, Correlation Coefficient,ROC Area, etc. (future)
  - Estimates of confidence intervals
- Displays for comparisons and run-time (future)
- Published results
  - For WFO users
  - For external customers
  - For collaboration

An Example of Local Verification



Figure(66): Example 2. Discrimination RMSE and ME for high and low stages: Forecasts, Calibration

National RVF Verification: Suggested End State

- Verify RFC issued stage time series forecasts
- Compute Actual and Persistence statistics
- Sort by
  - lead-time at daily time-steps
  - Above and below flood stage by observations
  - Fast, Medium and Slow responses
- Report RMSE skill score, POD, FAR, ME, MAE
- Collaborator access to electronic archives of forecasts and observations.

Local Probability Verification: End State

- All ensembles archived
  - Generated precipitation and temperature
  - With and without post-processor
- All input forecasts archived
- Climate as control forecast
- Sorting options similar to deterministic

Local Probability Verification: End State, cont ...

- Metrics
  - RPS, RPSS, PS (Brier Score)
  - ROC Diagrams, Area under the ROC curve
  - Ranked Histograms
  - Reliability Diagrams
  - Discrimination Diagrams
- Run time displays
- Displays to support comparisons
- Published results for both internal users and external customers.

Probabilistic Forecast Performance Measure

Statement of Need: The National Weather Service (NWS) has been producing probabilistic river forecasts since FY 2000. The NWS needs to implement a new performance measure to reflect the validity of its probabilistic river forecasts.

Performance Measure Guidance

- Meaningful to the audience
  - Internal and External to NWS
- Feasible
  - Data Availability
  - Data Collection, Management, Analysis and Reporting are Possible

Probabilistic Forecast Performance Measure

Probabilistic forecast reliability can be assessed using a Brier Score (*BS*):

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

where:

p_i is the forecast probability a flood will occur within the next x days for the ith forecast location or event, and

$o_i = 1$ if a flood occurred and $o_i = 0$ if a flood did not occur.

Probabilistic Forecast Performance Measure

Potential Implementation Steps:

- RFCs compute *x*-day probabilistic river forecasts and store this modeled output immediately.

- The NWS Archive Data Base Team will determine how the RFCs should store the probabilistic river forecasts (ensembles) and the forecast exceedance distributions. When the Archive Team has decided how the forecasts are to be stored, move the stored data to the appropriate location.

- Continue development of the ProbVS prototype to make it operationally ready. Update the software to get flood stage from archive DB, to add the Brier Score calculations, to generate the needed distributions and dump them out in XML, to read the distributions directly, to gather the observations and compute 0 or 1 from the archive DB, to dump out an XML file of the scores. A GUI will be necessary to control the data.

Verifying verification

- Success when
  - Actual verification metrics are used for decision making,
  - All enhancements are tied to expected improvements in specific verification metrics for specific groups of forecasts,
  - Success of enhancements is tracked through to the verification.

Useful References

- Forecast Verification, Ian T. Joliffe and David B. Stephenson ed., Wiley, 2003
- Statistical Methods in Atmospheric Sciences, Daniel S. Wilks, Academic Press, 1995
  - 2nd Edition en route
- Forecast Verification Web Page
  http://www.bom.gov.au/bmrc/wefor/staff/eee/verif/verif_web_page.html