



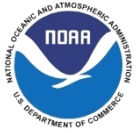
HYACINTS seminar: 09/10/09

**Quantitative evaluation of
probabilistic prediction systems
with observations**

Dr. James Brown

Hydrologic Ensemble Prediction Group, NOAA/NWS

james.d.brown@noaa.gov



Context

Probabilistic methods widely used

- Used in research, less in operations (e.g. PoP)
- Often based on 'ensembles' (Monte Carlo)
- Ensemble predictions are information rich

Evaluation is crucial (focus on obs.)

- Can we trust predictions? For what conditions?
- How can we best improve them (error sources)?
- Many methods (math., atmos., medical, econ. ,...)
- But, in hydrology, evaluation is patchy

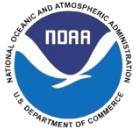


Contents

- 1. Distribution-oriented (DO) evaluation**
 - Joint distribution of predictions & observations
- 2. Attributes of quality**
 - These arise from joint distribution
- 3. Measures of quality (metrics)**
 - Vary in detail: lump all attributes or separate
- 4. Examples from real application**
 - NWS Ensemble Streamflow Prediction system



1. Distribution-oriented approaches



DO evaluation

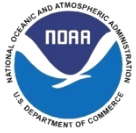
Predictions and observations

- Q , cont. (e.g. flow). We forecast and observe.
- Consider a discrete event (e.g. flood): $\{Q > q_v\}$.
- Prob. prediction (y) and observation (x).

$$y_i = \Pr[Q > q_v], \quad x_i = \{1 \text{ if } Q > q_v, \text{ else } 0\} \quad i = 1, \dots, n$$

How good is our model for $\{Q > q_v\}$?

- Two ways to look at joint distribution:
- $f(x, y) = a(x | y) \cdot b(y)$ “calibration-refinement”
- $f(x, y) = c(y | x) \cdot d(x)$ “likelihood-base-rate”



Considerations

What does $f(x,y)$ represent?

- Is Q variable in space and time? What support?
- What about forecast lead-time?

Can we estimate its properties?

- Pairs: time and/or space substitutes as repetition
- ...if max. daily flow, 1-day ahead, at Station A...
- ...over 1 year: $\{[x_1, y_1], \dots, [x_{365}, y_{365}]\}$
- In this case, pool data in time
- Try to avoid model calibration period (split?)



2. (Some) quality attributes



(Some) attributes of quality

Calibration-refinement: $a(x|y) \cdot b(y)$

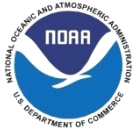
- Reliable if (e.g.): $E[x | y = p] = p \quad \forall p$
- “When $y = 0.2$, should observe 20% of time”
- Sharp if: $y \rightarrow 0$ or 1
- Aim: “maximize sharpness subject to reliability”

Likelihood-base-rate: $c(y|x) \cdot d(x)$

- Discriminatory if (e.g.):
 $E[y | x = 1] \gg E[y | x = 0]$
- “Forecasts easily separate flood from no flood”



3. (Some) quality metrics



(Some) quality metrics

1. Exploratory metrics (plot data pairs)

2. Lumped metrics ('scores')

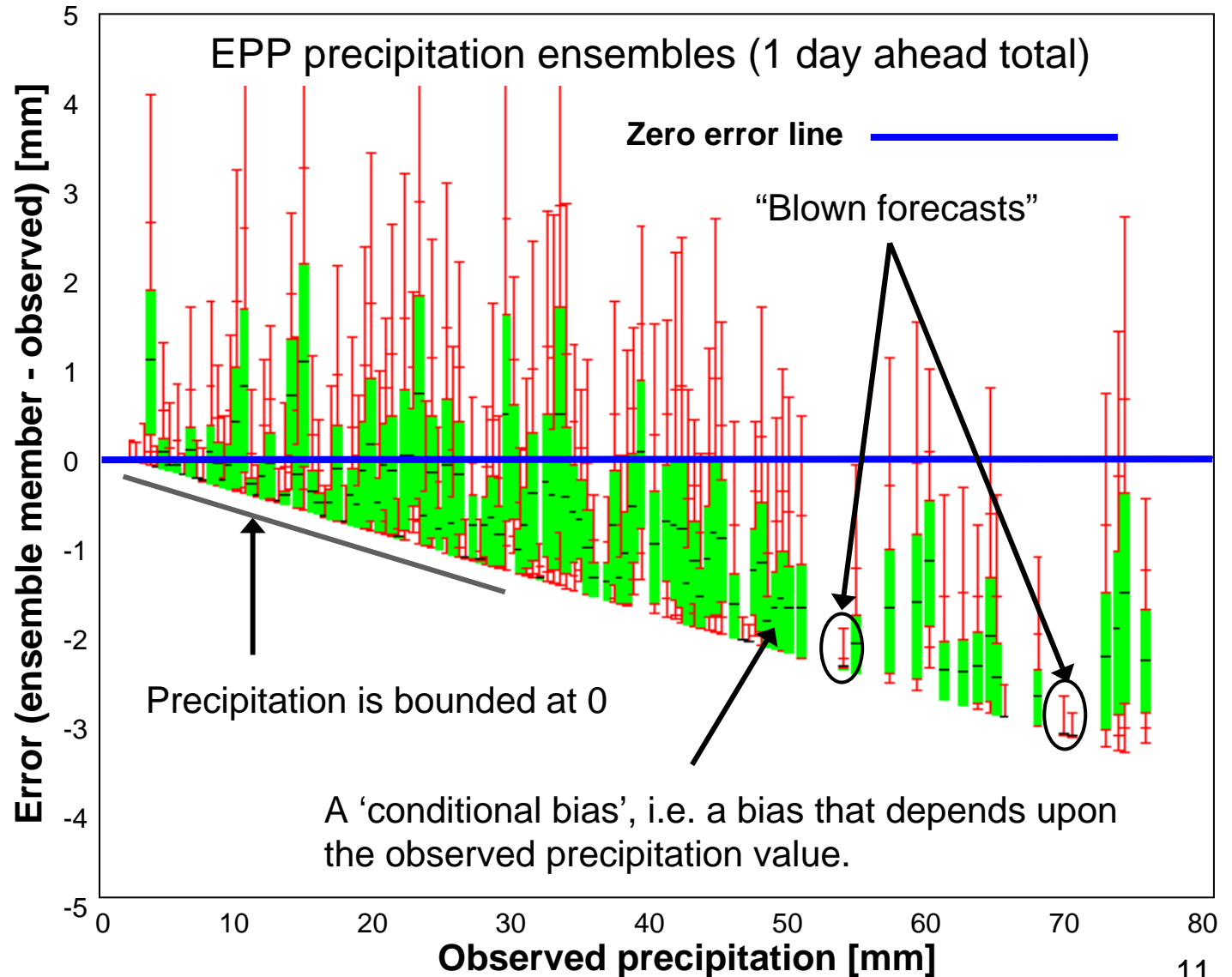
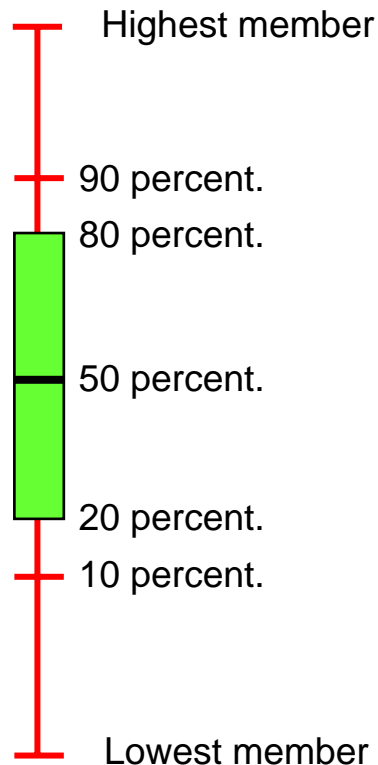
- Lumps all quality attributes (i.e. gross error)
- Often lumped over many events (e.g. $\forall q \in \mathfrak{R}$)
- Include skill scores (performance over baseline)

3. Attribute-specific metrics

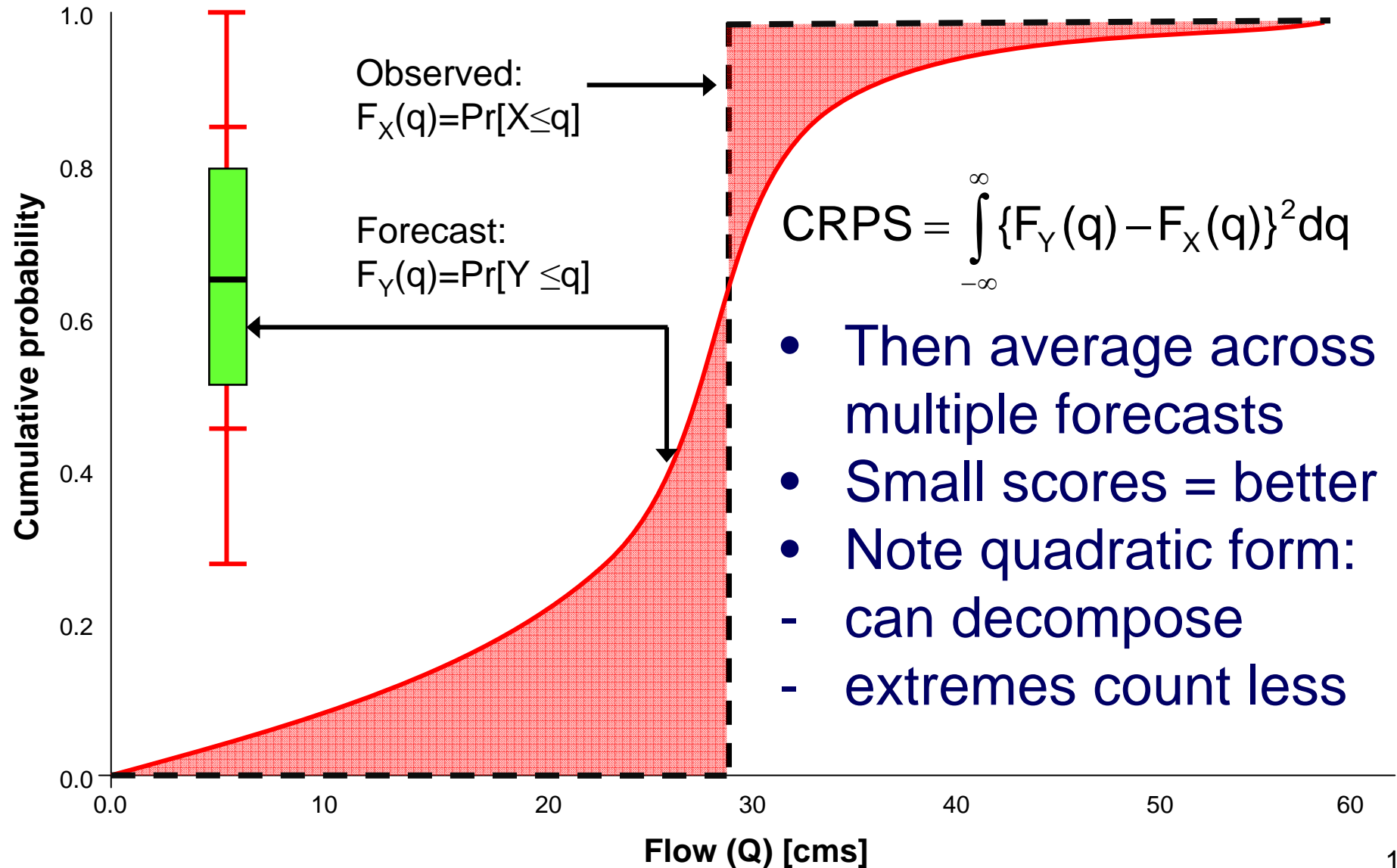
- Event reliability & sharpness (Reliability Diag.)
- Event discrimination (Relative Operating Char.)

Exploratory metric: box plots

'Error' for 1 forecast



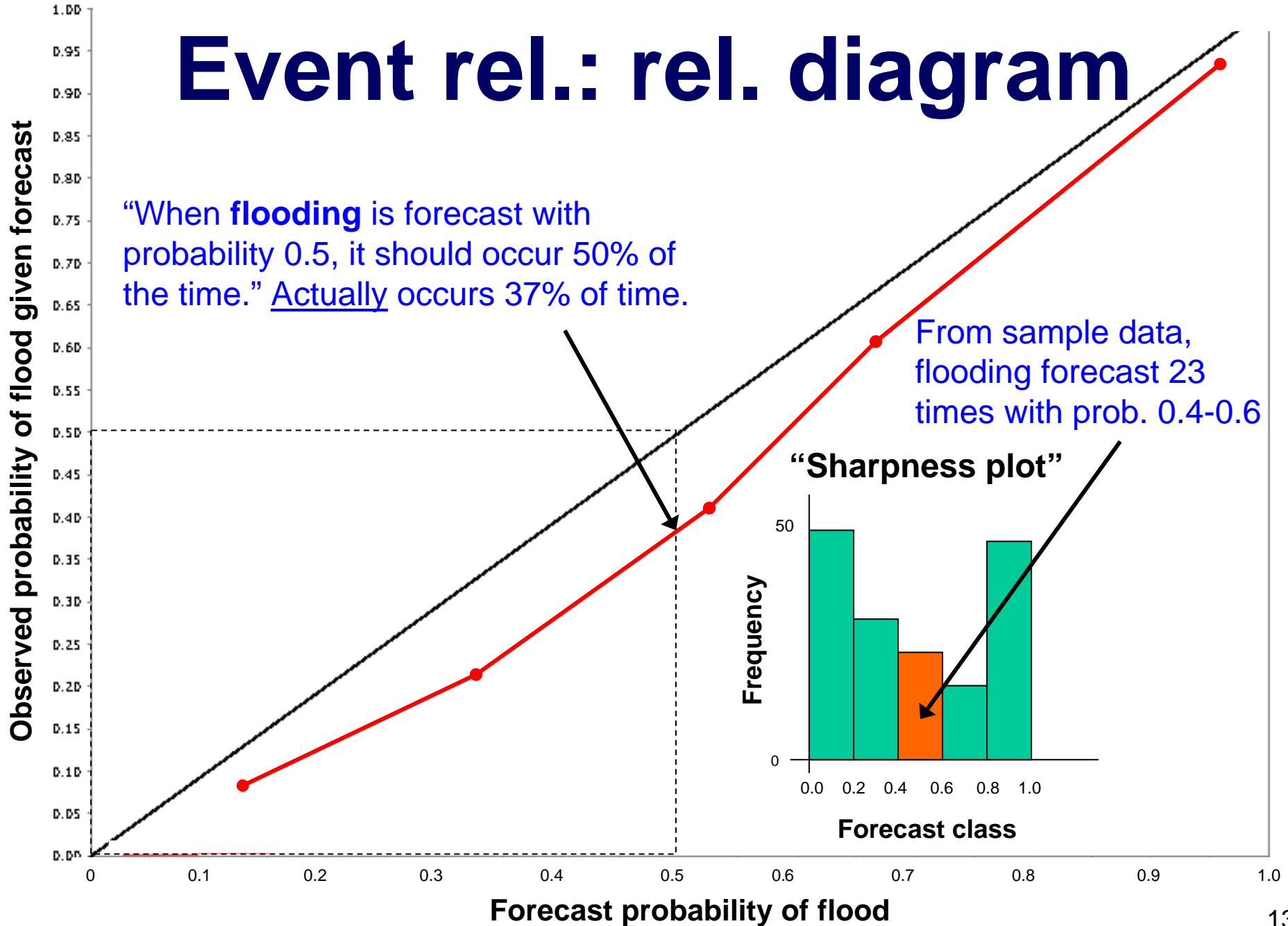
Lumped metric: Mean CRPS



Event rel.: rel. diagram

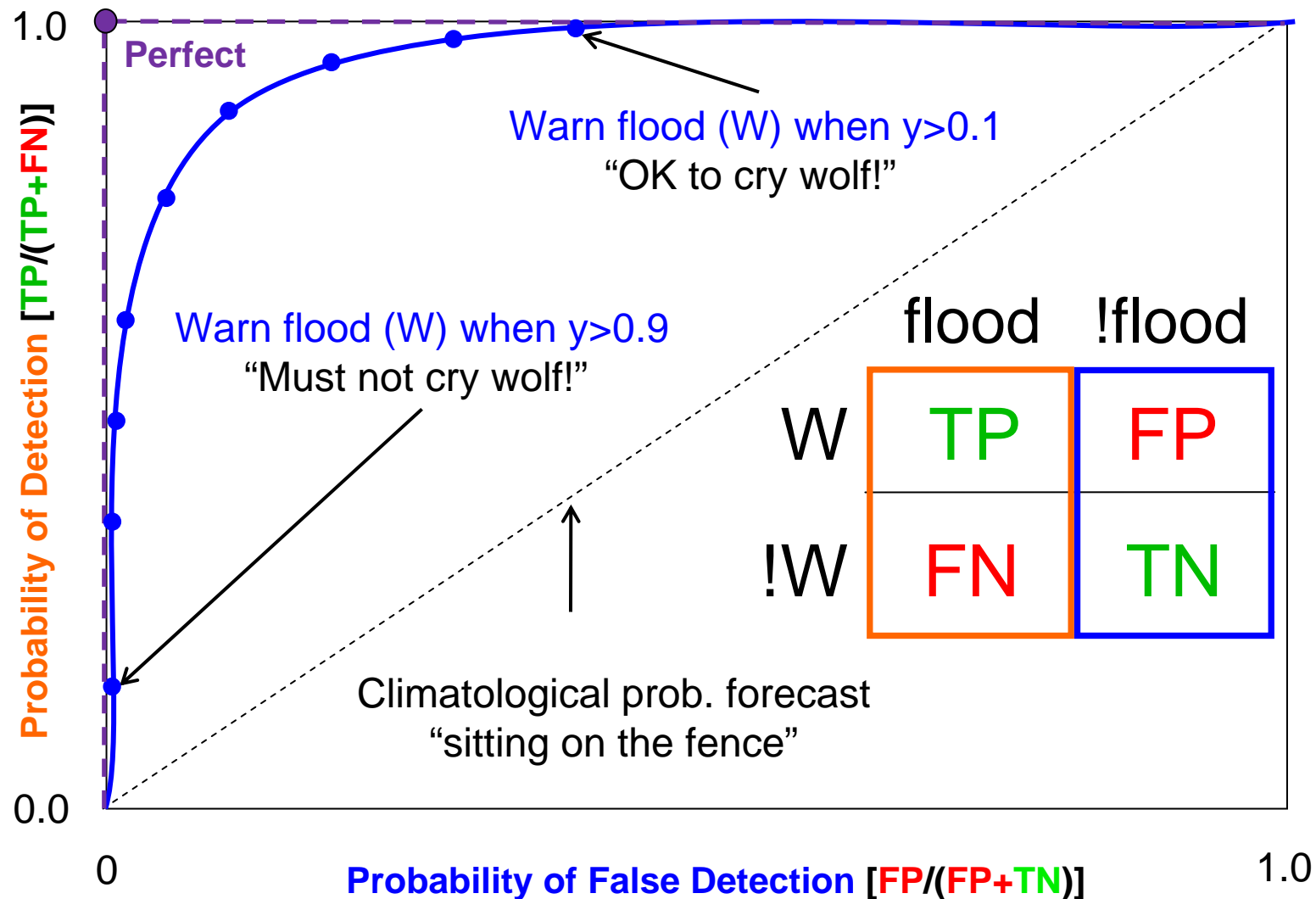
“When **flooding** is forecast with probability 0.5, it should occur 50% of the time.” Actually occurs 37% of time.

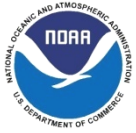
From sample data, flooding forecast 23 times with prob. 0.4-0.6





Event disc.: ROC (decision)

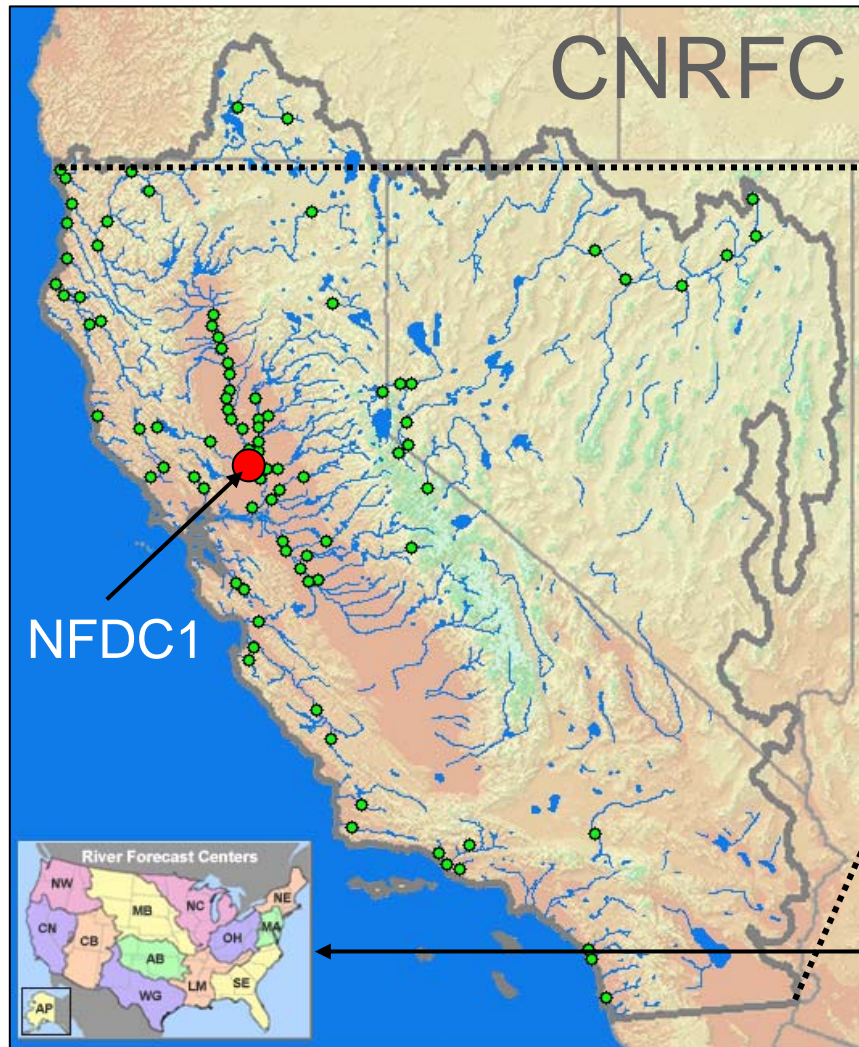




4. Real example



N. Fork, American (NFDC1)



NFDC1: dam inflow.
Lies on upslope of
Sierra Nevadas.

13 NWS River
Forecast Centers



Data available (NFDC1)

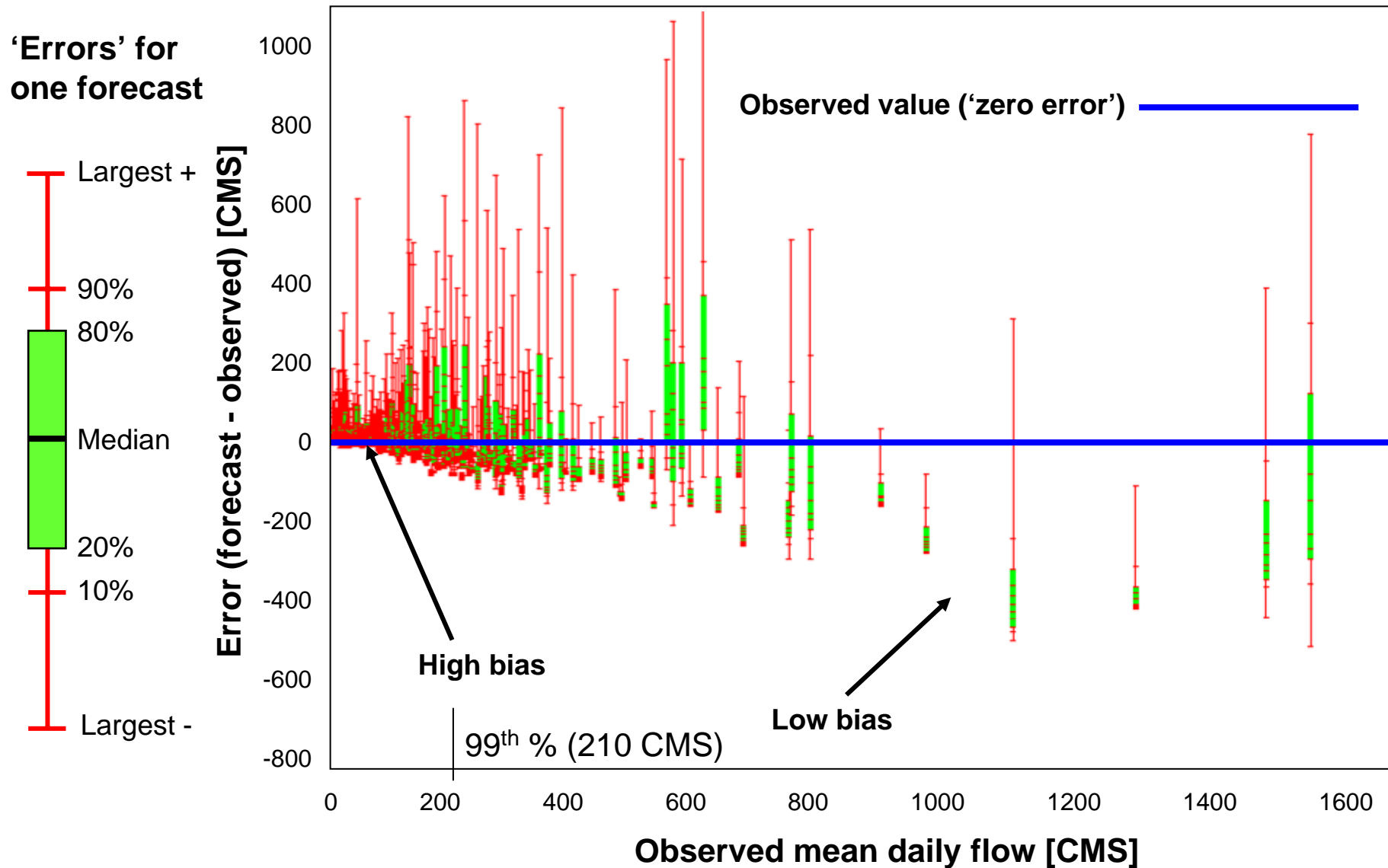
Streamflow ensemble forecasts

- Ensemble Streamflow Prediction system
- NWS RFS (SAC) w/ precip./temp. ensembles
- Hindcasts of mean daily flow 1979-2002
- Forecast lead times 1-14 days ahead
- NWSRFS is well-calibrated at NFDC1

Observed daily flows

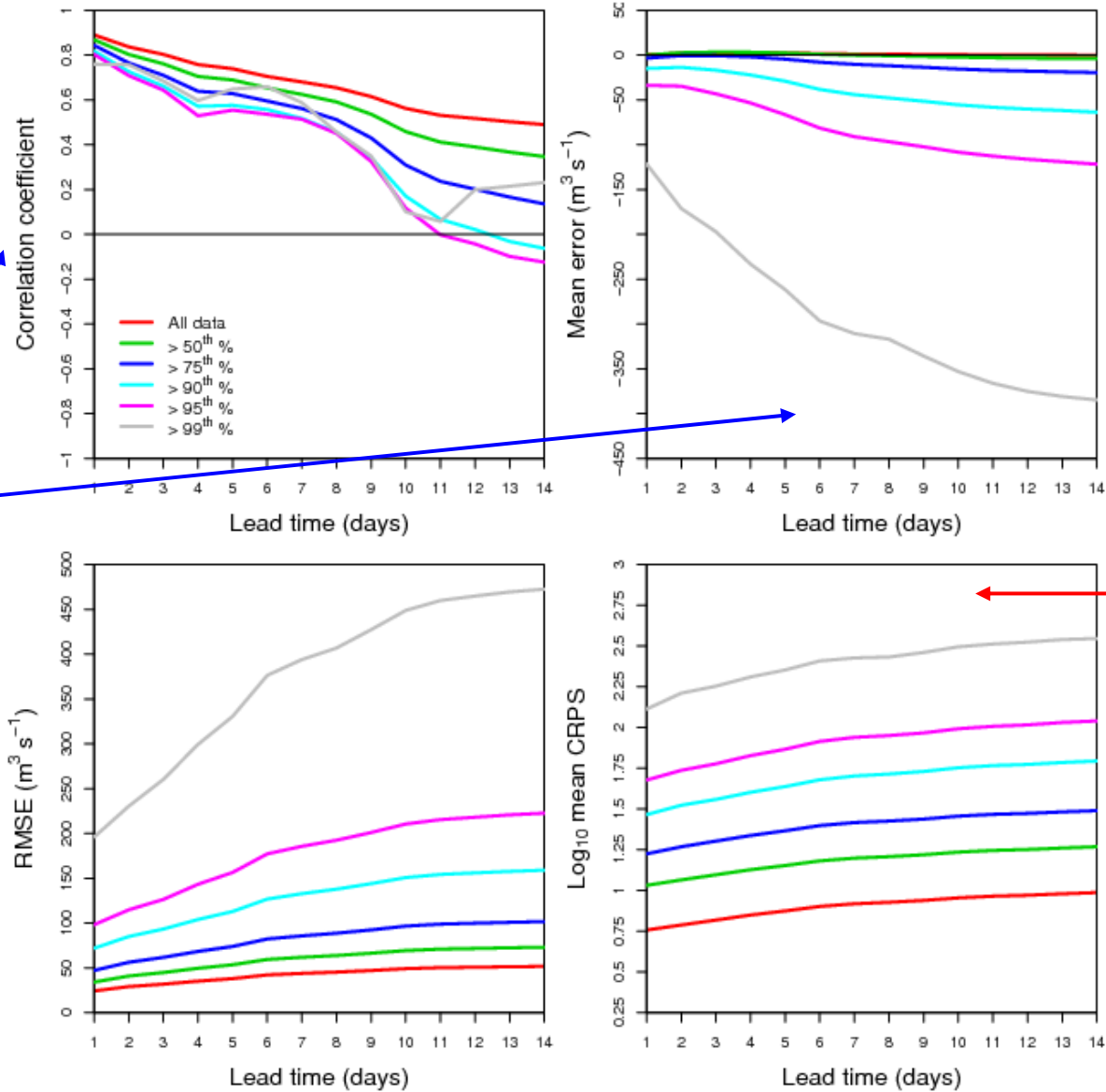
- USGS 6-hourly observed stage
- Convert to discharge and average for 1 day

Box plot of flow errors (day 1)



Lumped error statistics

Tests of ensemble mean

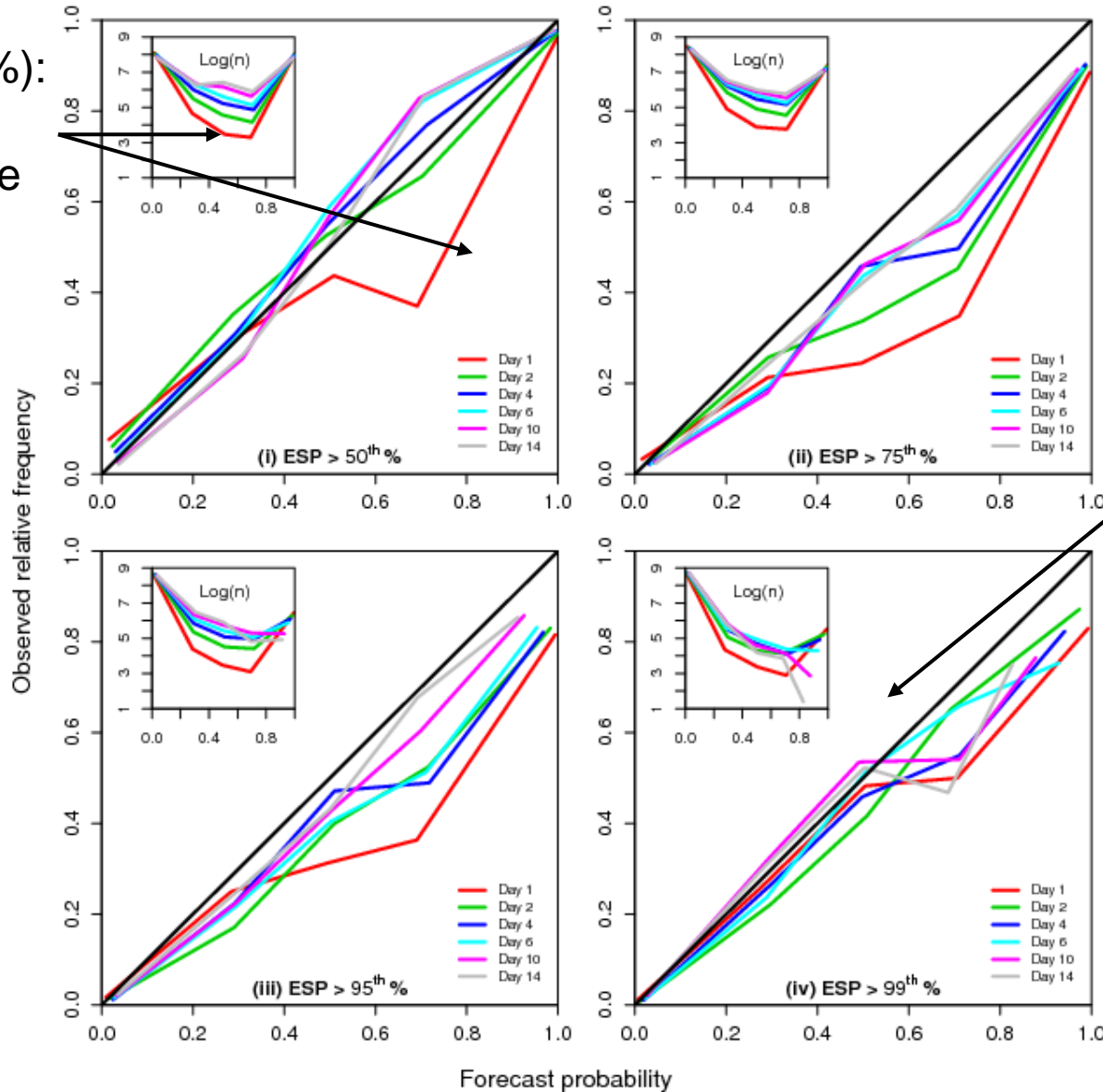


Lumped error in probability

Reliability

Day 1 (>50th%):
sharp, but a
little unreliable
(contrast
day 14).

No initial
condition
uncertainty
(all forcing).



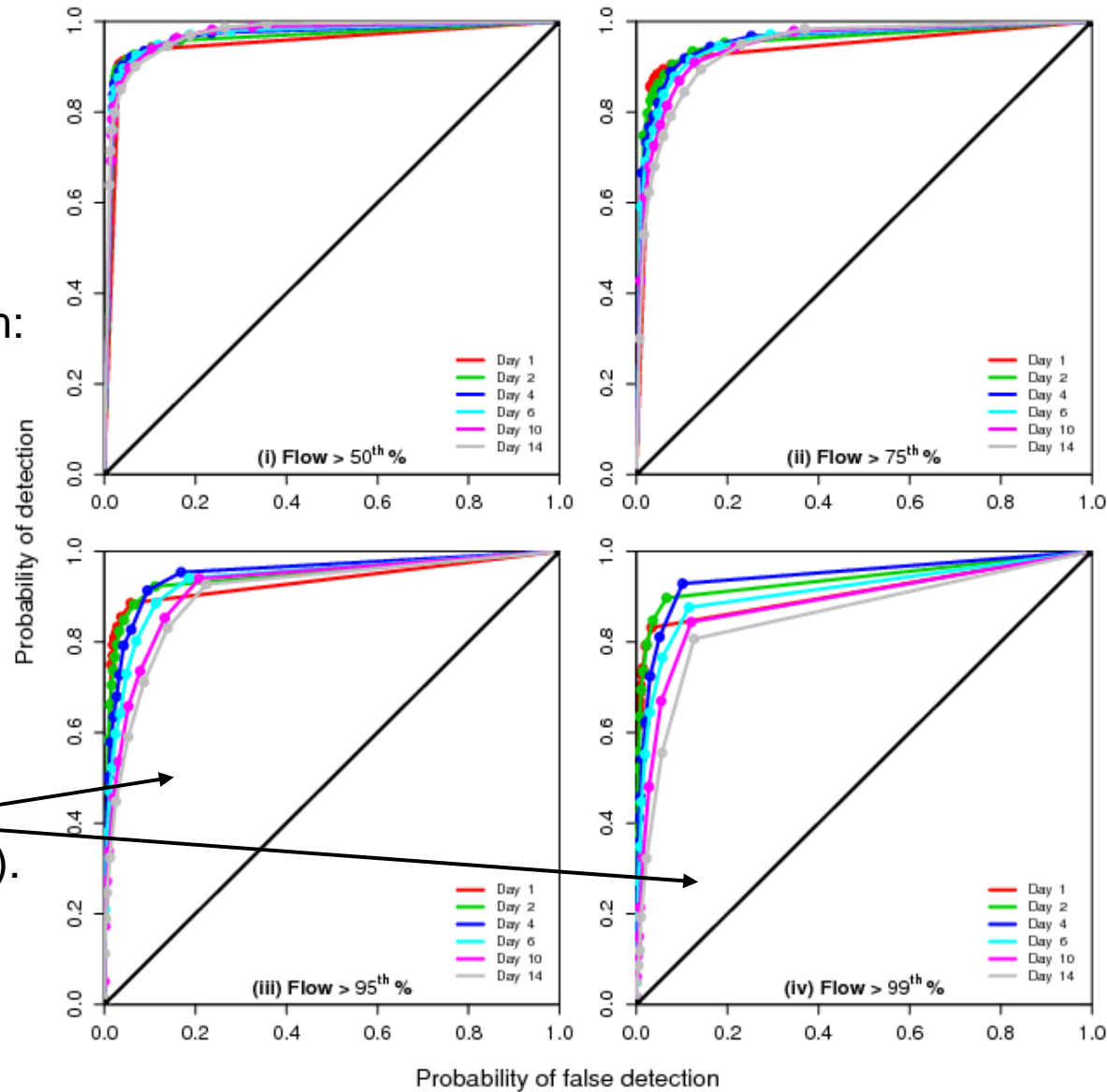
Day 14 (>99th%):
even though
cond. bias in
mean, probs.
are reasonably
reliable: due to
99% = 210 CMS
and spread is
good.

But, note
sample size.

ROC

Forecast easily beat climatology represented by diagonal (climatology system: forecast prob. is long-term average frequency)

Forcing very good for large events (orographic lifting = predictable).



Verification metrics to compute

Metrics to compute

Name	Property verified	Include?
Correlation coefficient	Ensemble mean	<input checked="" type="checkbox"/>
Mean error	Ensemble mean	<input checked="" type="checkbox"/>
Root mean squared error	Ensemble mean	<input checked="" type="checkbox"/>
Brier score	Ensemble distribution	<input checked="" type="checkbox"/>
Mean continuous ranked probability score	Ensemble distribution	<input checked="" type="checkbox"/>
Mean error of probability diagram	Ensemble distribution	<input checked="" type="checkbox"/>
Mean capture rate diagram	Ensemble distribution	<input checked="" type="checkbox"/>
Modified box plot pooled by lead time	Ensemble distribution	<input checked="" type="checkbox"/>
Modified box plot per lead time by observed value	Ensemble distribution	<input checked="" type="checkbox"/>
Relative operating characteristic	Ensemble distribution	<input checked="" type="checkbox"/>
Relative operating characteristic score	Ensemble distribution	<input checked="" type="checkbox"/>

Explanation of metric 'Mean continuous ranked probability score'

MEAN CONTINUOUS RANKED PROBABILITY SCORE (CRPS)

The CRPS summarizes the quality of a continuous probability forecast with a single number (a score). It measures the integrated squared difference between the cumulative distribution function (cdf) of a forecast, $F_Y(y)$ and the corresponding cdf of the observation, $\mathbf{1}\{\}$:

$$CRPS(x, F_Y) = \int_{-\infty}^{\infty} (F_Y(y) - \mathbf{1}(y \geq x))^2 dy$$

where $\mathbf{1}\{\}$ is a step function that reaches probability 1.0 for values greater than or equal to the observation, and has probability 0.0 otherwise. In practice, the CRPS is averaged across a number, n , of paired forecasts and observations, which leads to the mean CRPS:

$$\overline{CRPS} = 1/n \sum CRPS(x_i, F_{Y_i})$$

Parameters of metric 'Mean continuous ranked probability score'

Edit thresholds [optional]

Threshold values
All data
0.0
0.05
0.1

Add Delete

“Ensemble Verification System”

Public download soon at:

http://www.nws.noaa.gov/iao/iao_hydroSoftDoc.php

More

Save Run All

Back Next

Literature

- Bradley, A. A., Schwartz, S. S. and Hashino, T., 2004: Distributions-Oriented Verification of Ensemble Streamflow Predictions. *Journal of Hydrometeorology*, **5(3)**, 532-545.
- Brown, J.D., Demargne, J., Liu, Y. and Seo, D-J (submitted) The Ensemble Verification System (EVS): a software tool for verifying ensemble forecasts of hydrometeorological and hydrologic variables at discrete locations. Submitted to *Environmental Modelling and Software*. 52pp.
- Gneiting, T., F. Balabdaoui, and Raftery, A. E., 2007: Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, **69(2)**, 243 – 268.
- Hsu, W.-R. and Murphy, A.H., 1986: The attributes diagram: A geometrical framework for assessing the quality of probability forecasts. *International Journal of Forecasting*, **2**, 285-293.
- Jolliffe, I.T. and Stephenson, D.B. (eds), 2003: *Forecast Verification: A Practitioner's Guide in Atmospheric Science*. Chichester: John Wiley and Sons, 240pp.
- Mason, S.J. and Graham N.E., 2002: Areas beneath the relative operating characteristics (ROC) and relative operating levels (ROL) curves: Statistical significance and interpretation, *Quarterly Journal of the Royal Meteorological Society*, **30**, 291-303.
- Murphy, A. H. and Winkler, R.L., 1987: A general framework for forecast verification. *Monthly Weather Review*, **115**, 1330-1338.
- Wilks, D.S., 2006: *Statistical Methods in the Atmospheric Sciences*, 2nd ed. Academic Press, 627pp.