National DOH Workshop
07/16/08

A brief refresher on verifying ensemble forecasts

James Brown

James.D.Brown@noaa.gov

1. Types of ensemble Verification metric

Aim: reduce forecast bias

Many types of bias. For example:

- **Over- or under-forecasting (e.g. ensemble mean consistently too low or high).**
- **Too little uncertainty in ensemble forecast ("underspread").**
- **Bias that increases under specific conditions, ("conditional bias").**
- **Bias resulting from poor model physics ("unreliable") or resolution ("unresolved").**

Types of metrics

Many types of metrics

- Reflects many different types of bias
- Four-dimensions reviewed here

1. Deterministic vs. ensemble approach

- Convert ensemble forecast to single-valued forecast by choosing “best guess” (mean).
- Apply single-valued metrics (RMSE etc.)
- Easy to understand, but inadequate.

Types of metrics

2. Absolute vs. relative quality

a) Absolute: metric for one forecast model

b) Relative: *skill* of one model over another

Skill needs a metric and reference

3. Detailed vs. summarized

- Detailed visualization of pairs (e.g. box plots)
- …to 'one-number' scores (e.g. mean CRPS).
- Both valuable (even for one application).

Types of metrics

4. Reliability vs. discrimination

- **When Y was forecast, what was observed?**

  "Our model predicts a 90% chance of flooding."

  RELIABLE if observed 9/10 times issued.

- **When X was observed, what was forecast?**

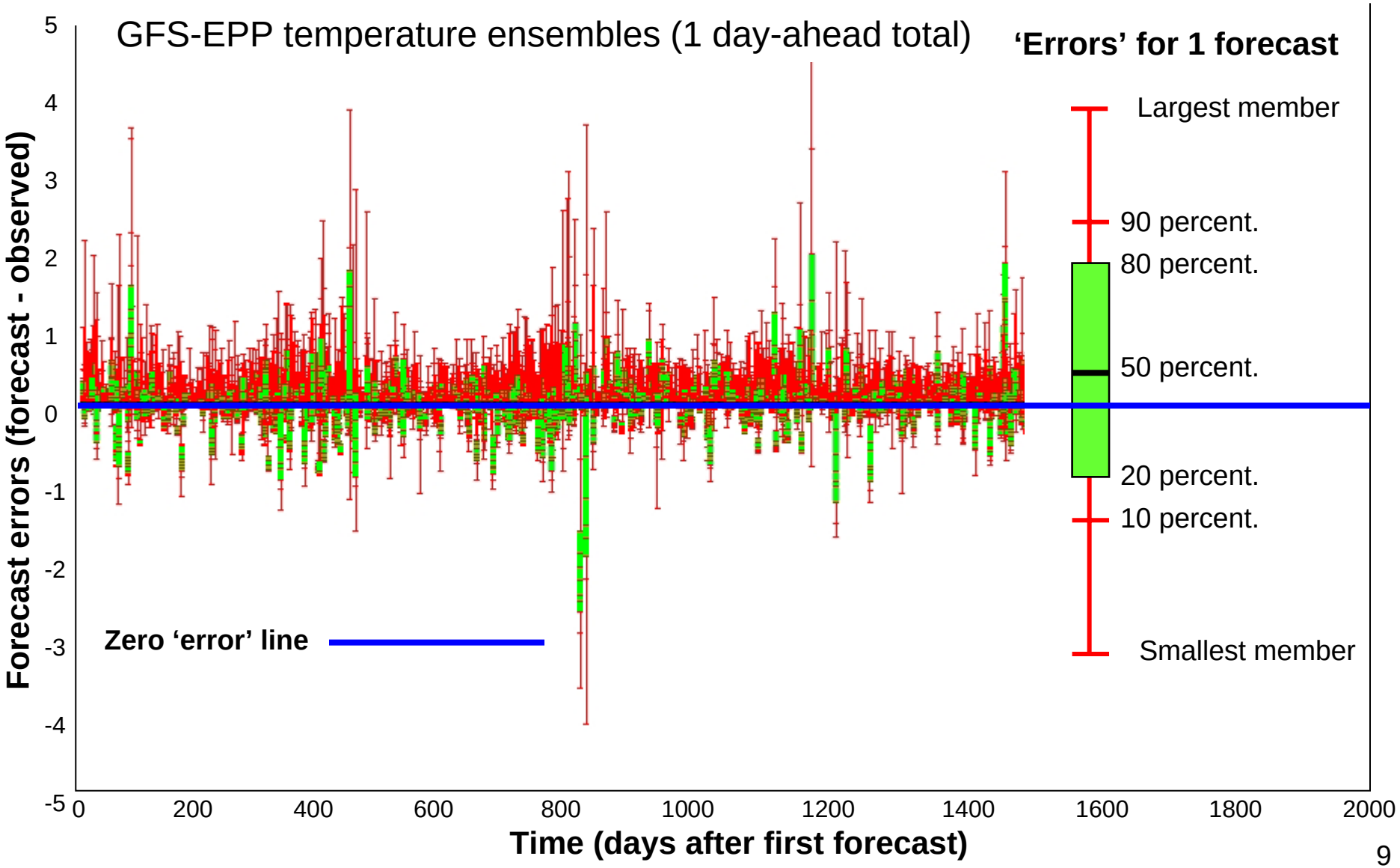
  "When we observe Action Stage only, our model predicts a 100% chance of Flood Stage."

  Cannot DISCRIMINATE between AS and FS.

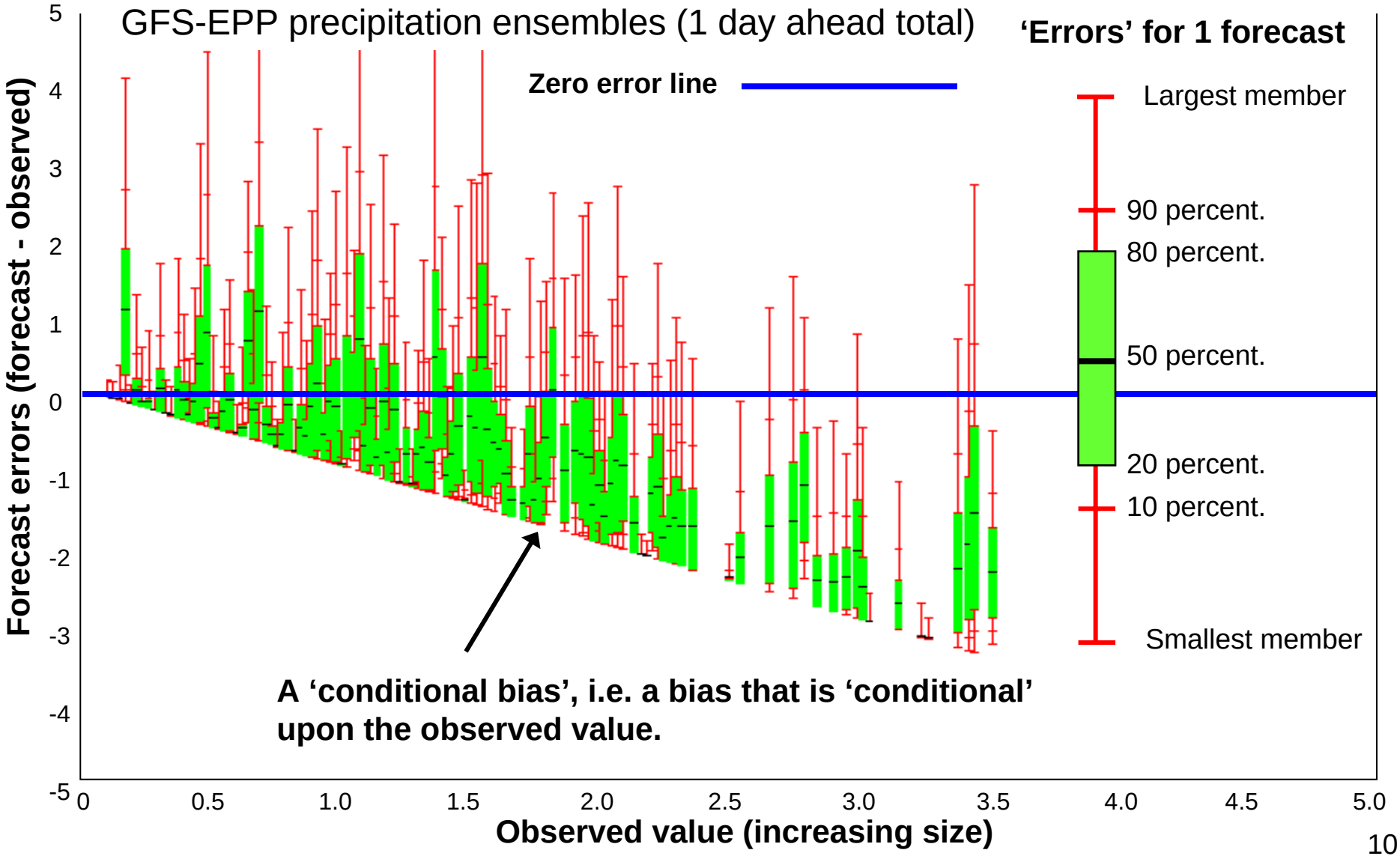
2. Examples of key metrics you will see today and how they are calculated

Detailed vs. summarized

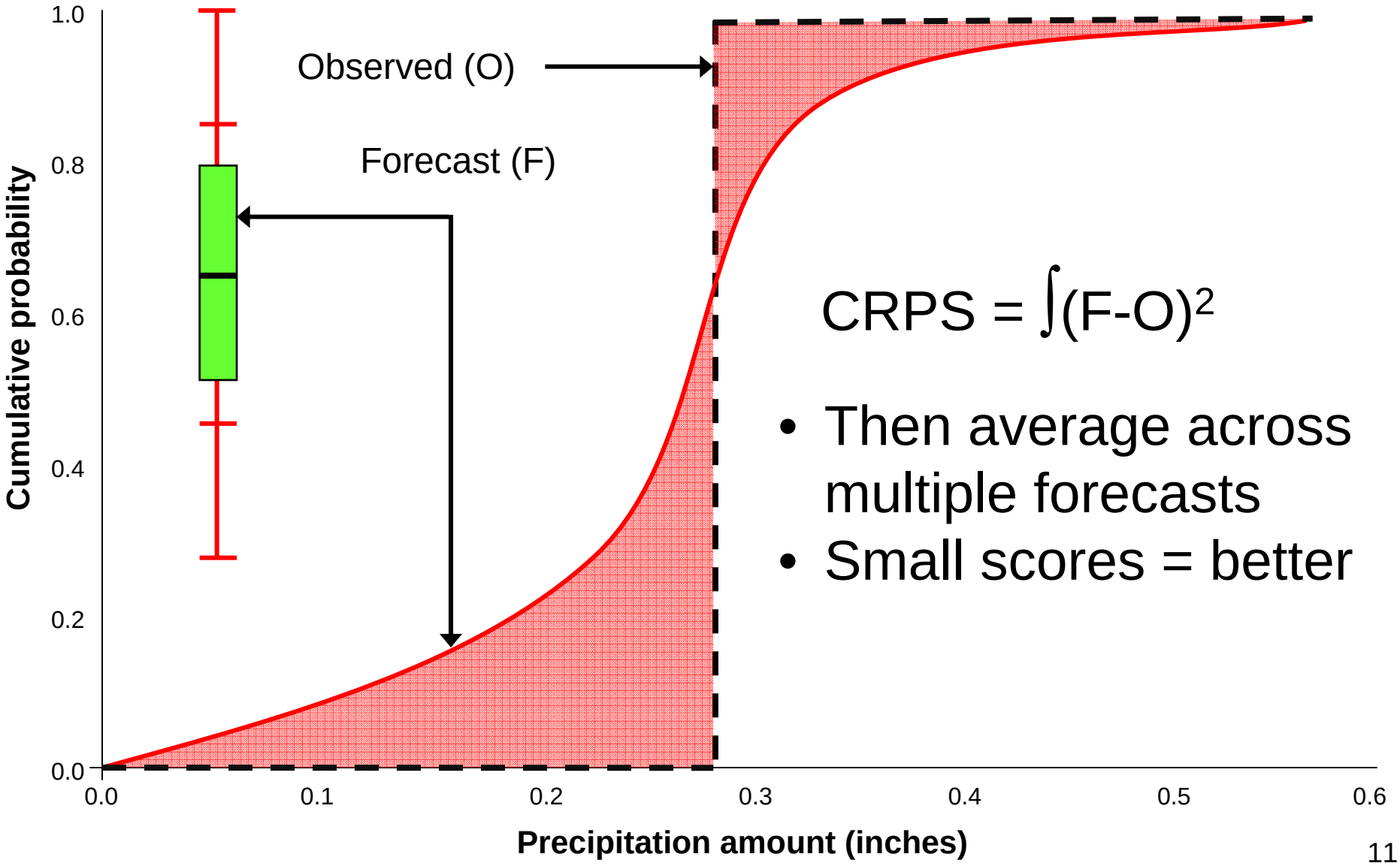
Most detailed (box plot)



Most detailed (box plot)



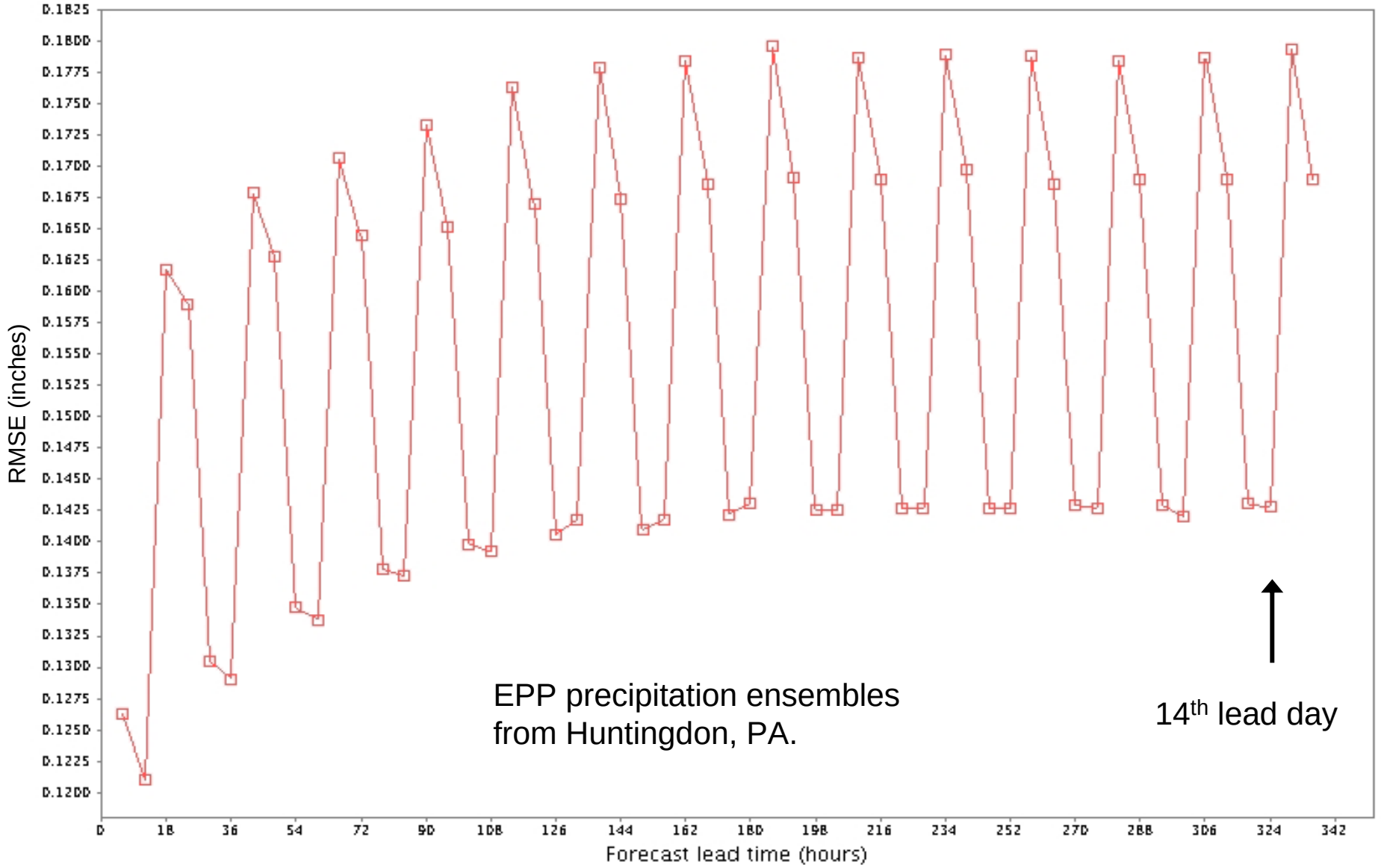
CRPS (much less detailed)



$$CRPS = \int (F-O)^2$$

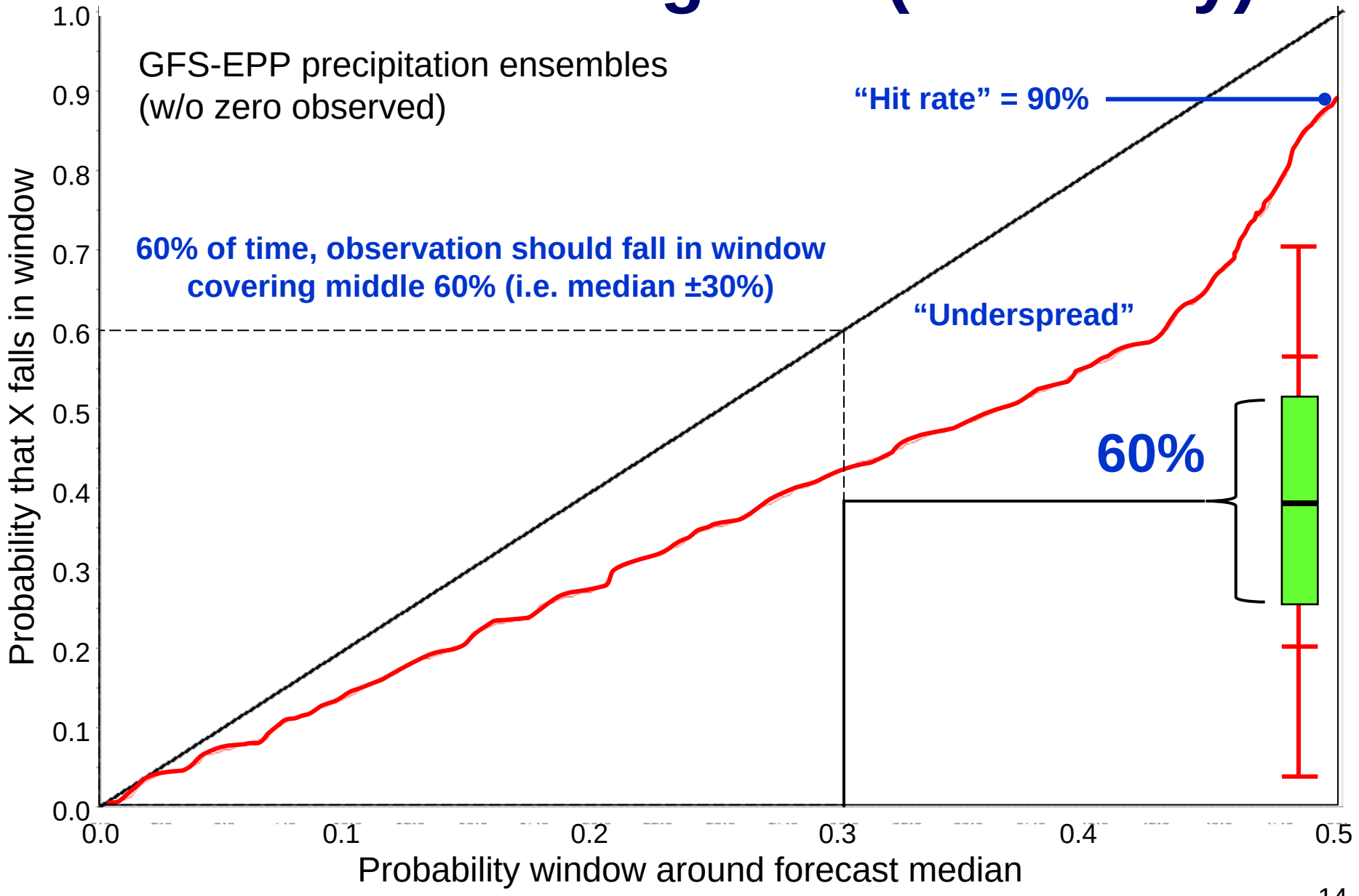
- Then average across multiple forecasts
- Small scores = better

RMSE of mean (least detailed)



Reliability vs. discrimination

Cumulative Talagrand (reliability)

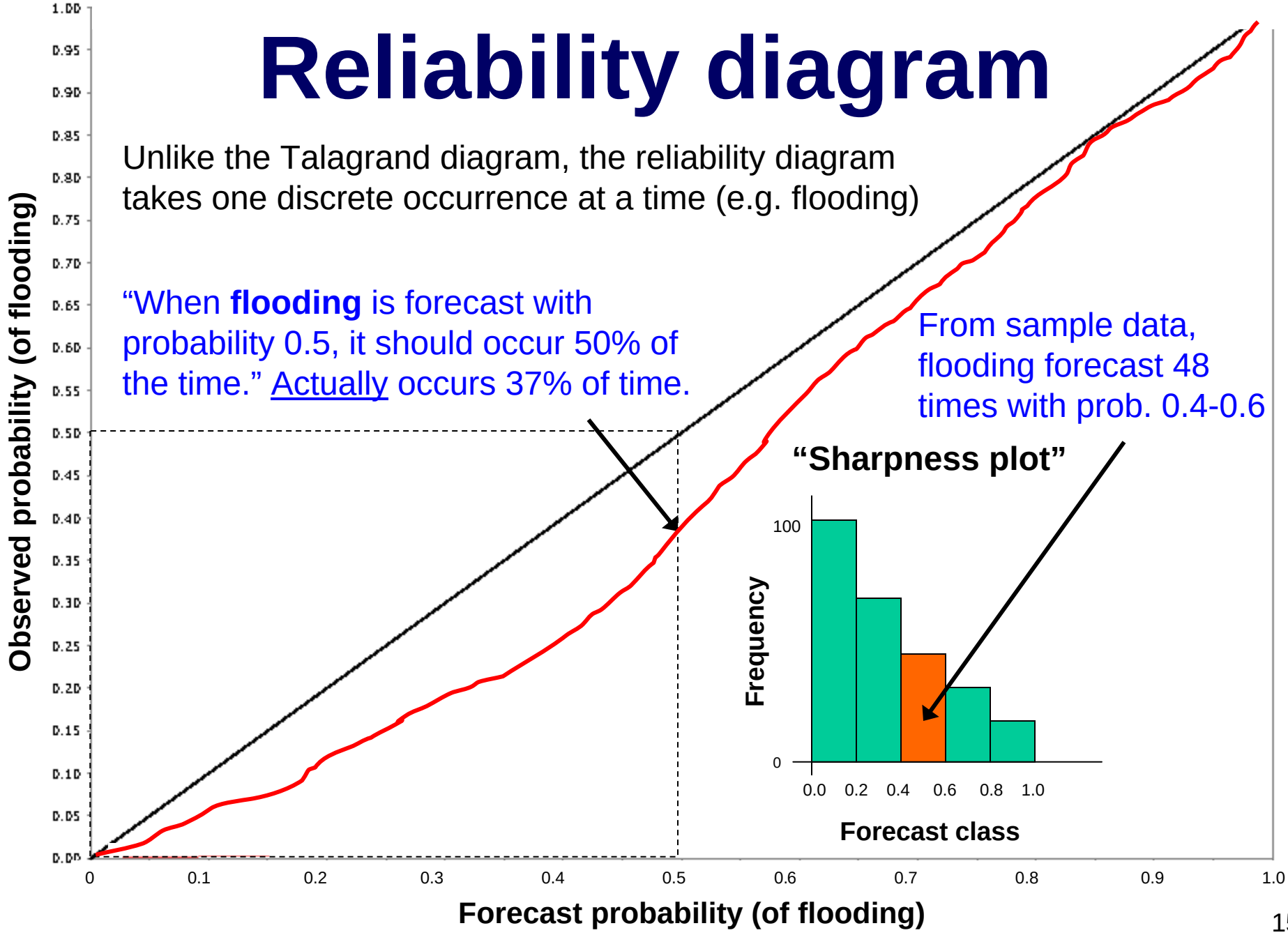


Reliability diagram

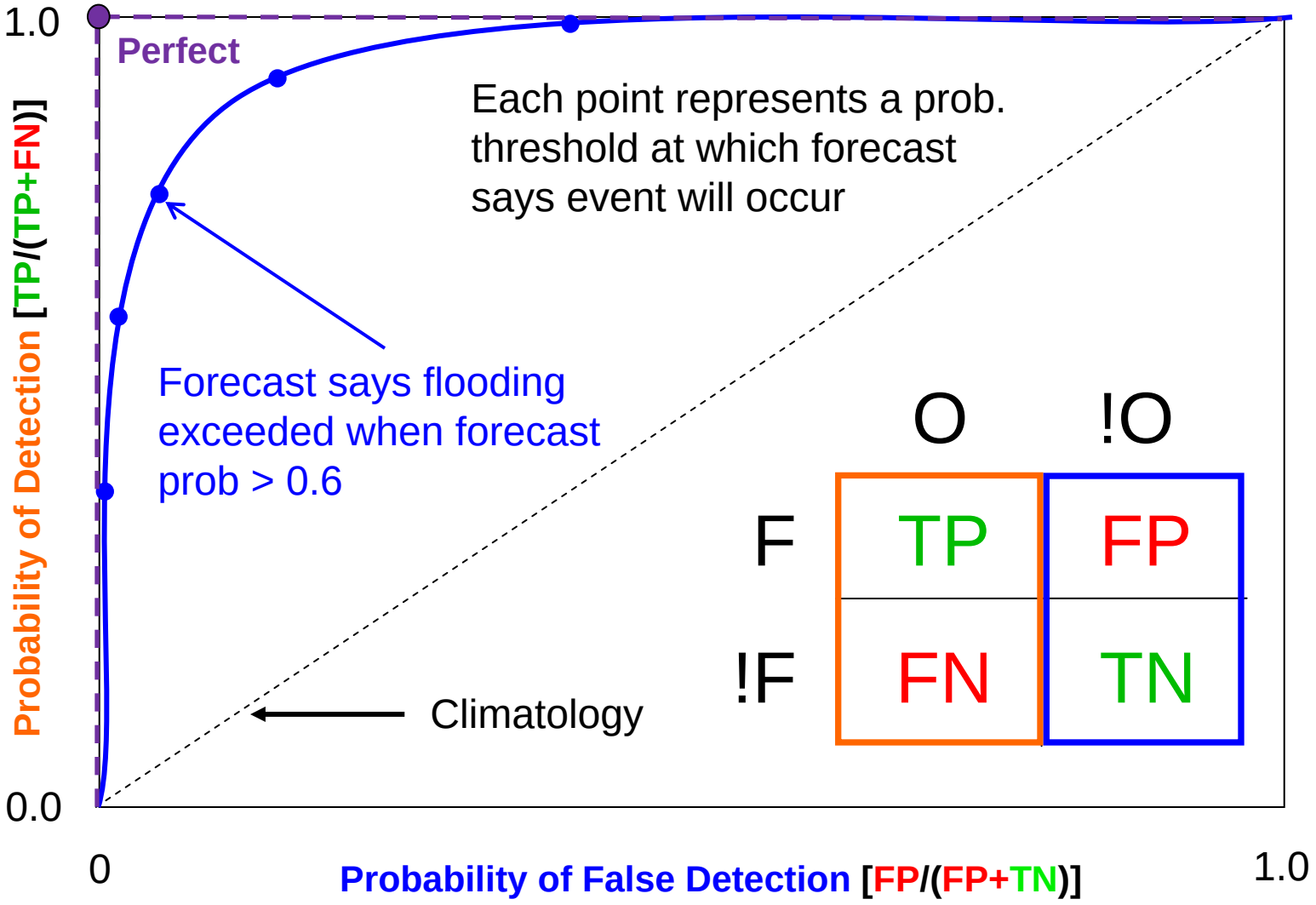
Unlike the Talagrand diagram, the reliability diagram takes one discrete occurrence at a time (e.g. flooding)

"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 48 times with prob. 0.4-0.6



ROC plot (discrimination)



Questions ???

Extra slides (for questions)

Why verify?

Many scientific and applied reasons

- E.g. “Completing the Forecast”

Separating bias from noise

- Forecasts will never be “error-free”
- Aim: to minimize *consistent* errors (bias)
- Eventually, just left with random noise
- XEFS/HEFS aims to do this
- Verification is needed to identify bias

How can we verify?

Collect past forecasts and observations

- Database of past forecasts and observations.
- Pair every forecast with its associated obs.
- Does the pairing make sense?

Then examine their joint statistics

- Cannot identify *bias* from a single pair.
- And we only *sample* the "true" relationship.
- Hence, we rely on statistics.

How can we verify?

Clearly, we make some assumptions….

“Stationarity” (increases sample size)

- **Ensembles: many-to-one pairing (many members vs. one observation)**
- **We collect together many pairs and assume each forecast is one realization of a *stationary* process (= many-to-many pairing).**
- **Does *not* imply identical forecasts, but some statistical properties must be constant.**

How can we verify?

Assumption of discrete events

- **Ensembles give us probabilities of events.**
- **Continuous distributions have *infinitely* many events. How to deal with this?**

a) ….sometimes, interested in events that are *inherently* discrete (e.g. flood: [stage > FS]).

b) ….some metrics integrate over all events (e.g. Mean CRPS).

c) ….otherwise, we must simply use thresholds.