# Metrics for probabilities

## Many ways to classify metrics

1. **Tests for single-valued property (e.g. mean)**
2. **Tests of broader forecast distribution**

- **Both may involve reference forecasts (“skill”)**

## Caveats in testing probabilities

- **Observed probabilities require many events**
- **Big assumption 1: we can ‘pool’ events**
- **Big assumption 2: observations are ‘good’**

# Continuous prob. forecasts

## Discrete/categorical forecasts

- Many metrics rely on discrete forecasts
- e.g. will it rain? {yes/no} (rain > 0.01)
- e.g. will it flood? {yes/no} (stage > flood level)

## What about continuous forecasts?

- An infinite number of events
- Arbitrary event thresholds (i.e. ‘bins’)?
- Typically, yes (and choice will affect results)

# Metrics vary by design

## Observation-centered metrics (discrim.)

- “What do forecasts do when observed do X”?
- i.e. “binning” in terms of observed
- e.g. Relative Operating Characteristic

## Forecast-centered metrics (reliability)

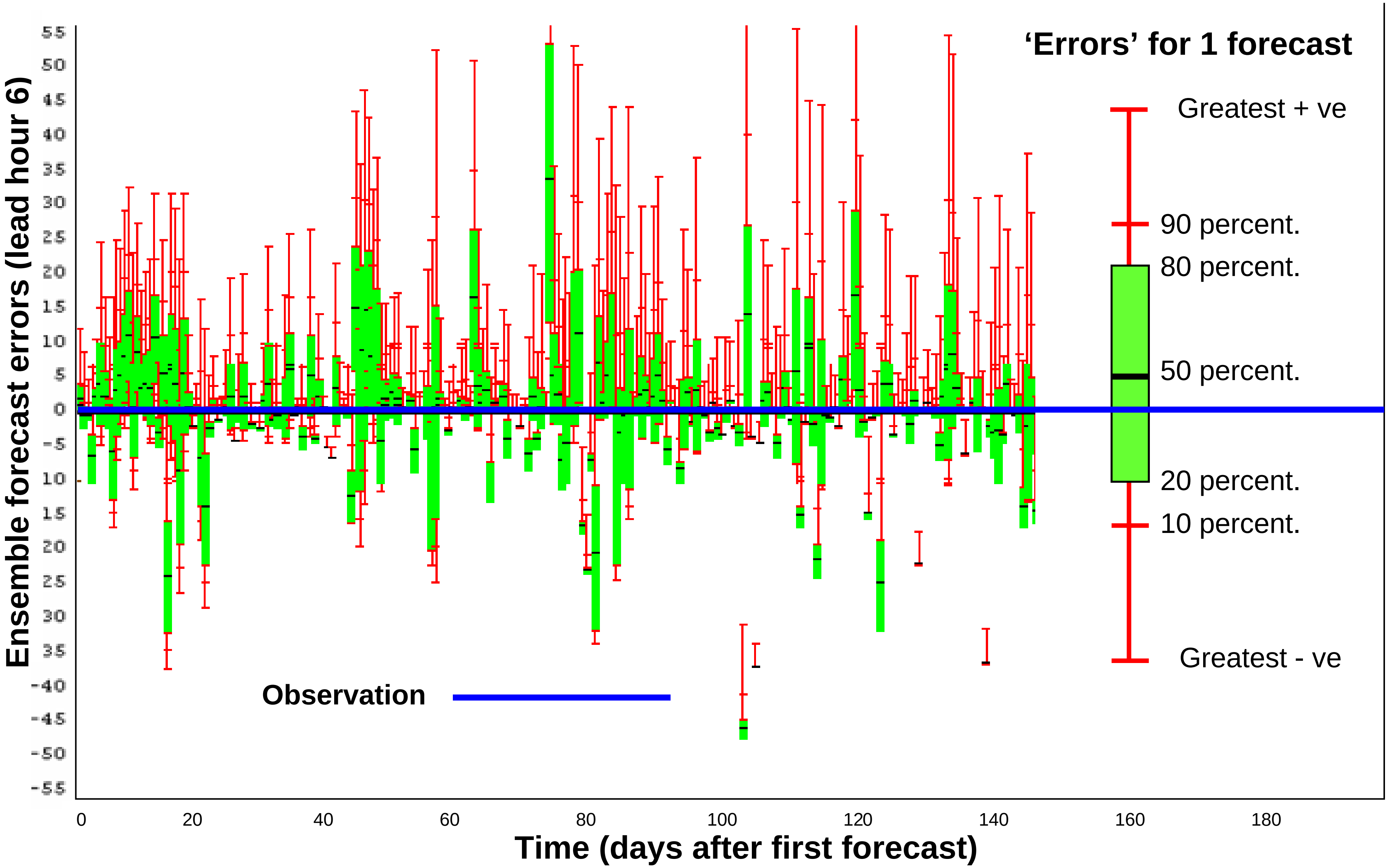
- “What do observed do when forecasts do Y”?
- i.e. “binning” in terms of forecasts
- e.g. Reliability Diagram

# Metrics vary in detail

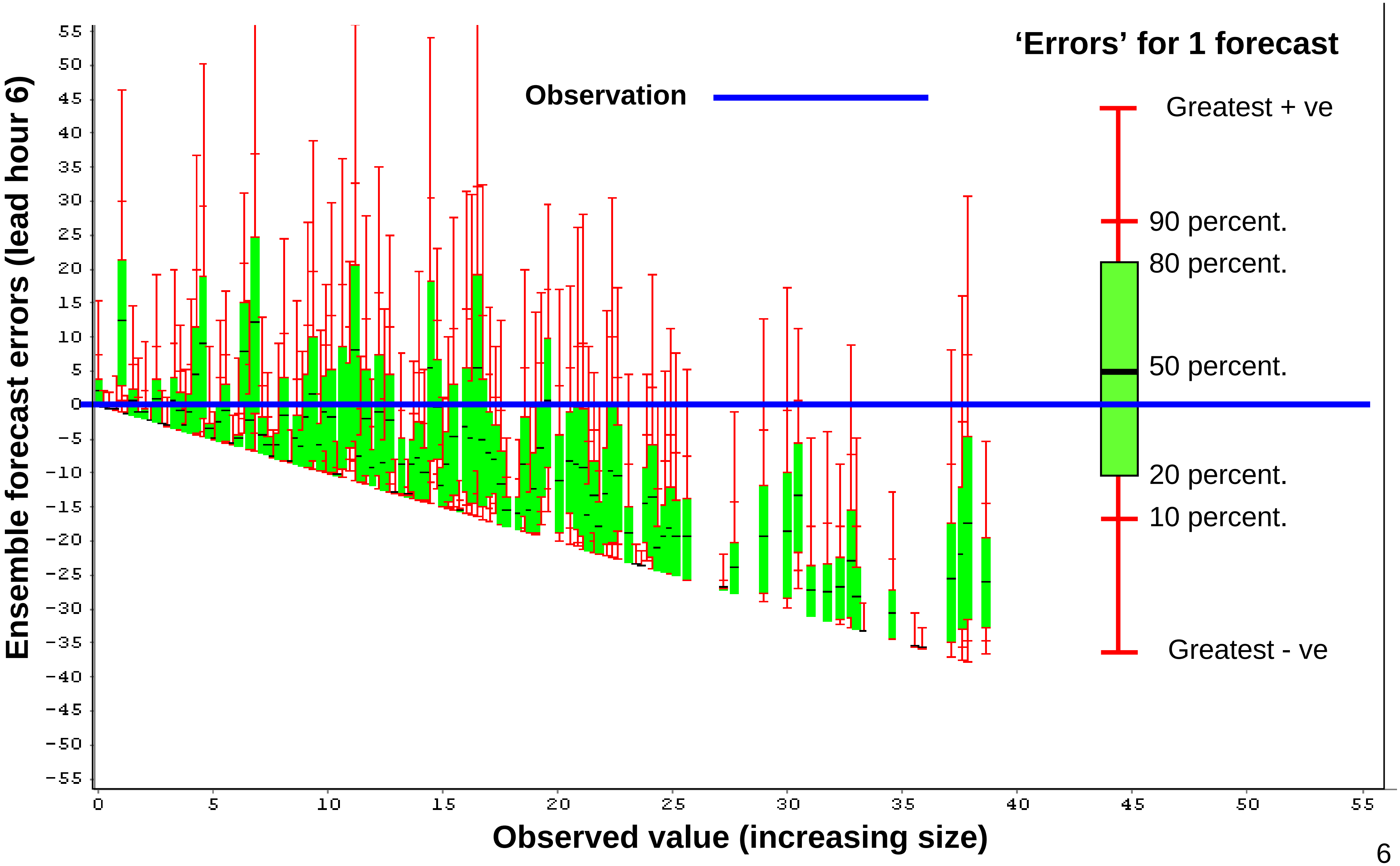
## Detail varies with verification question

- e.g. inspection of ‘blown’ forecasts (detailed)
- e.g. avg. reliability of flood forecast (< detail)
- e.g. rapid screening of forecasts (<< detail)

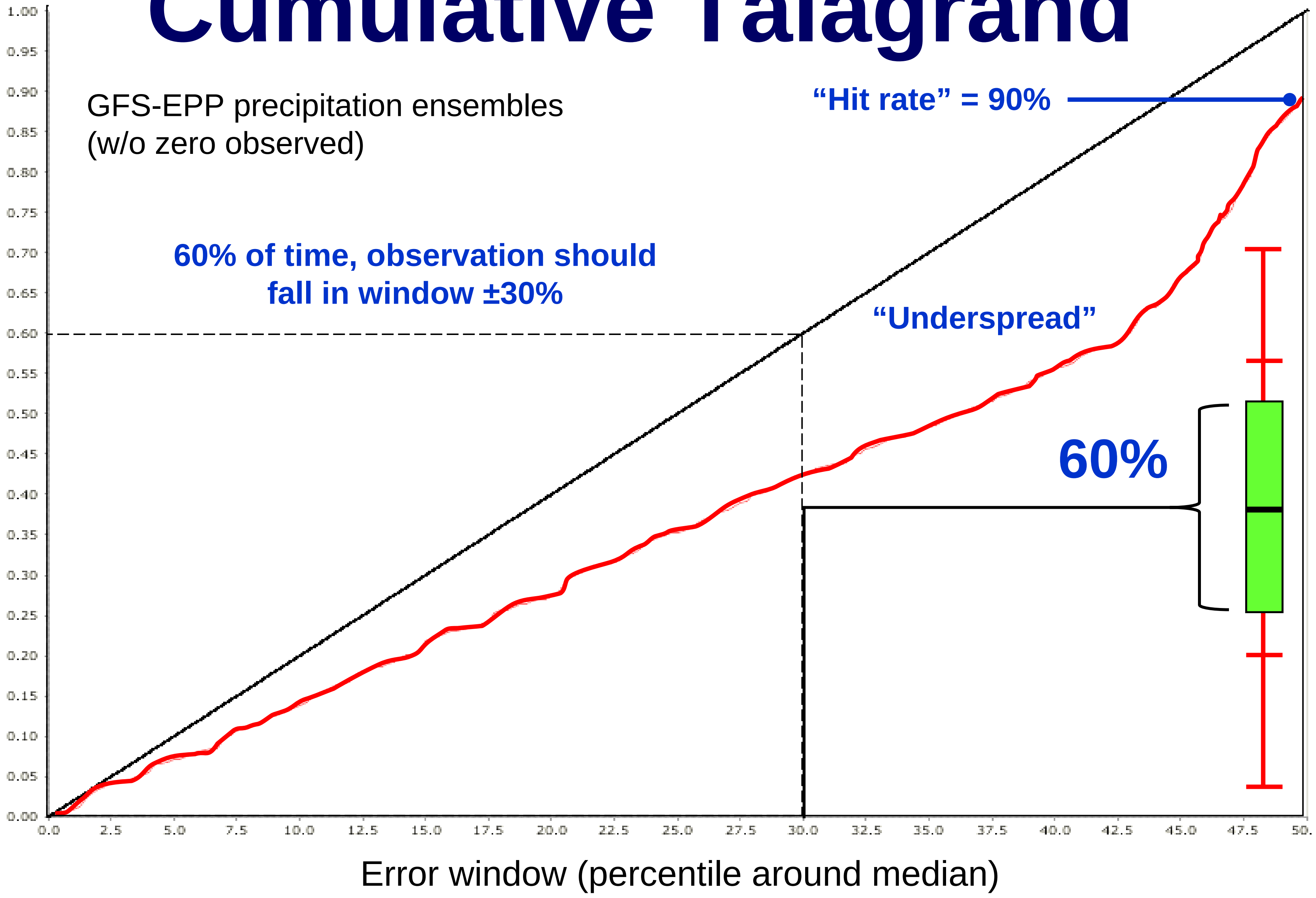
# Most detailed (box plot)



# Most detailed (box plot)



# Cumulative Talagrand



GFS-EPP precipitation ensembles
(w/o zero observed)

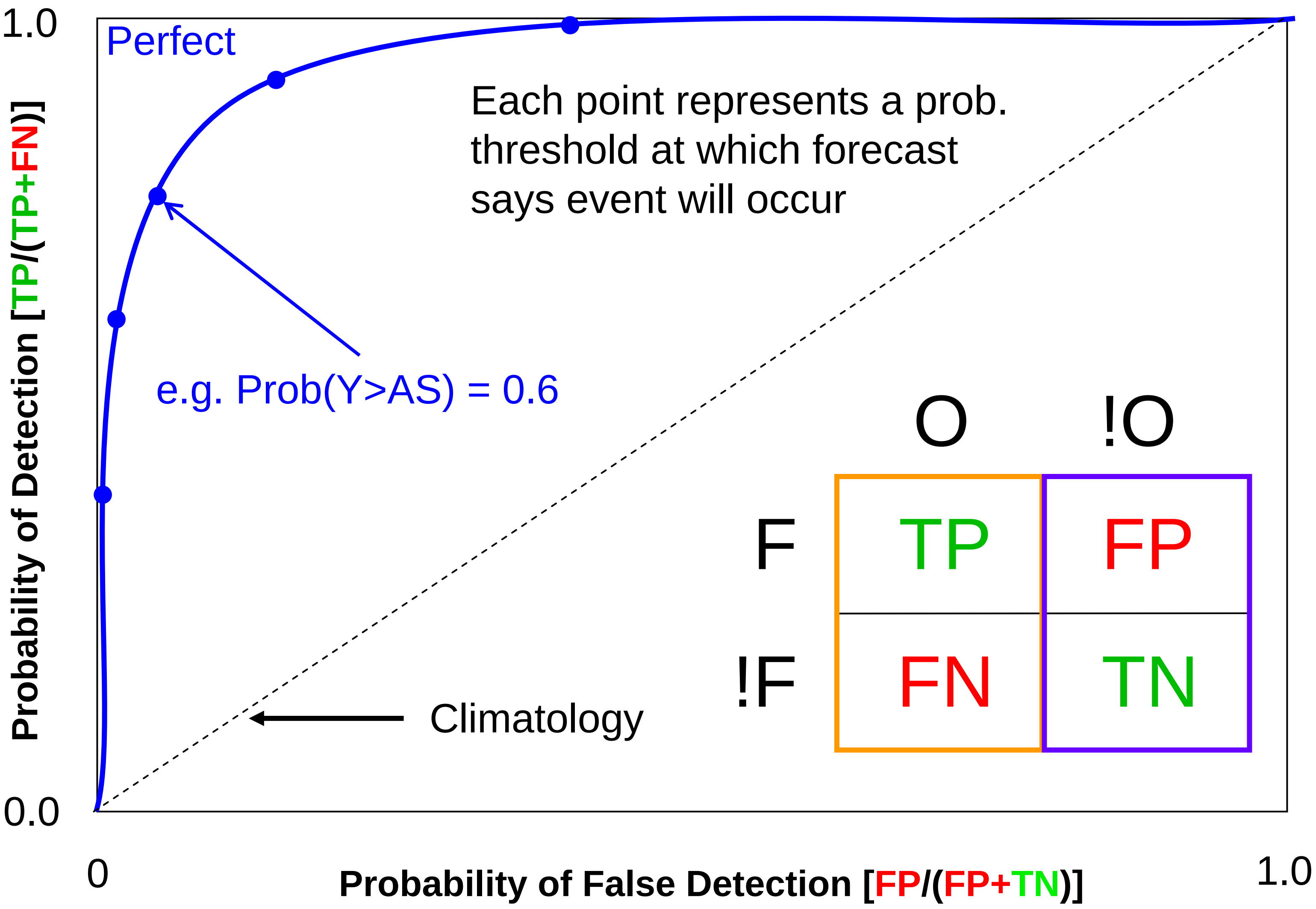
60% of time, observation should
fall in window ±30%

“Hit rate” = 90%

“Underspread”

60%

# ROC at Flood Action Stage

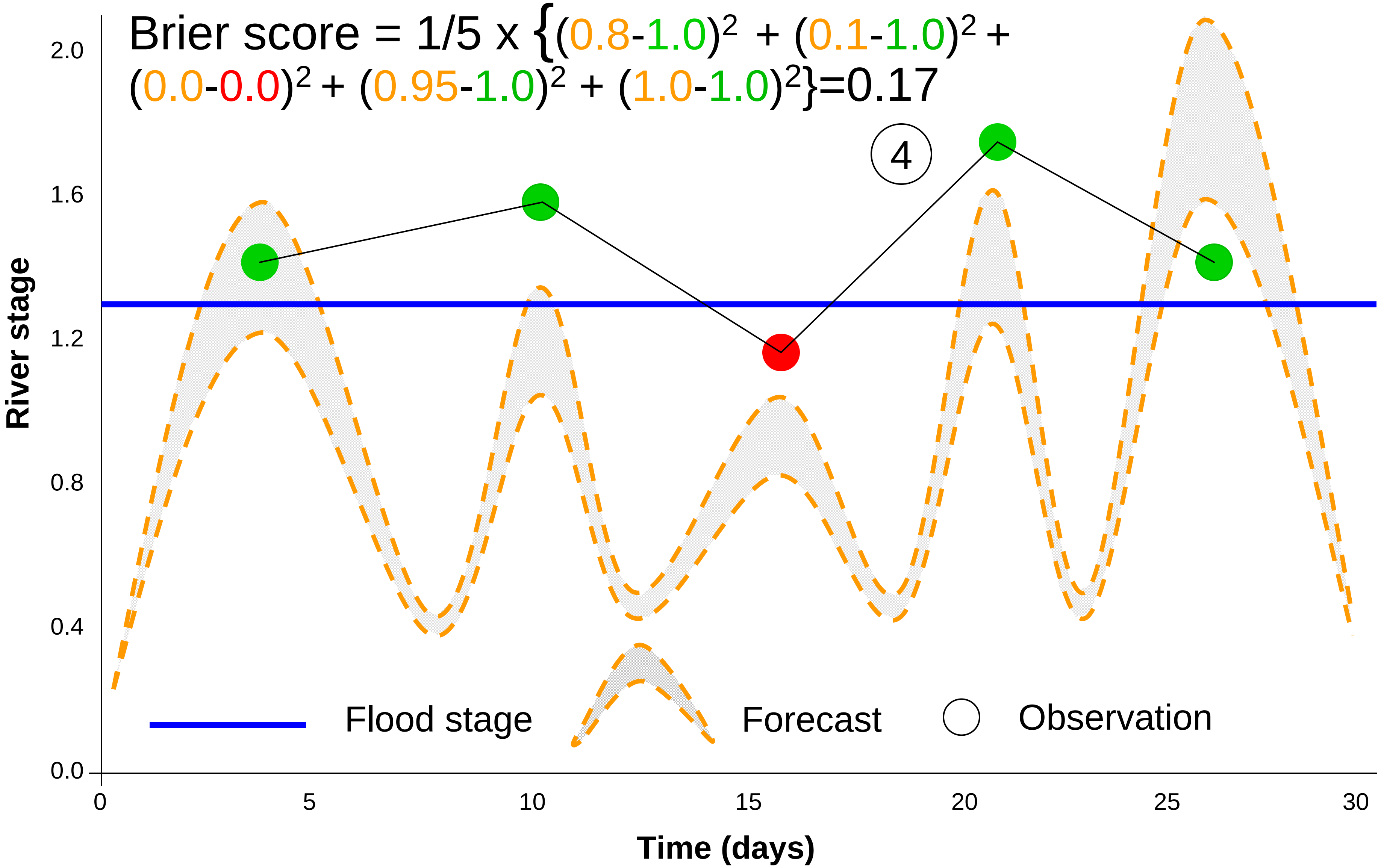


Each point represents a prob. threshold at which forecast says event will occur

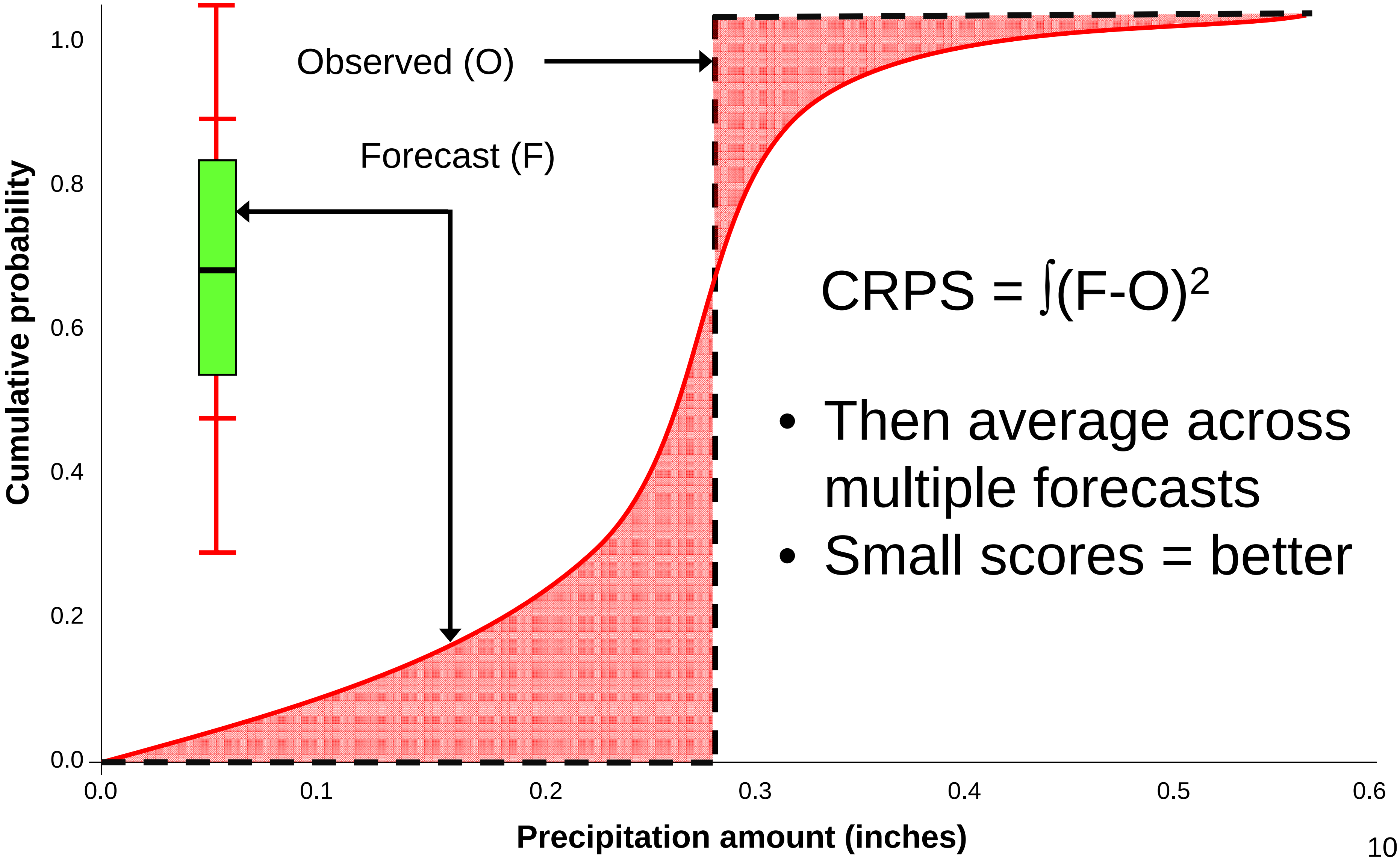
|  | O | !O |
|---|---|---|
| F | TP | FP |
| !F | FN | TN |

# Least detailed (a score)

$$\text{Brier score} = 1/5 \times \{(0.8-1.0)^2 + (0.1-1.0)^2 + (0.0-0.0)^2 + (0.95-1.0)^2 + (1.0-1.0)^2\}=0.17$$



# Least detailed (a score)



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

- Then average across multiple forecasts
- Small scores = better