

Concepts in verification

Matt Pocernich
pocernic@ucar.edu

Finley Tornado Data (1884)



LIEUTENANT JOHN F. FINLEY, SIGNAL CORPS, UNITED STATES ARMY.

John F. Finley

		Observed		
		Yes	No	Total
Forecast				
	Yes	28	72	100
	No	23	2680	2703
	Total	51	2752	2803

A success?

		Observed		
Forecast		Yes	No	Total
	Yes	28	72	100
	No	23	2680	2703
	Total	51	2752	2803

•Percent Correct = $(28+2680)/2803 = 96.6\% !!!!$

Maybe not.

		Observed		
Forecast		Yes	No	Total
	Yes	0	0	0
	No	51	2752	2803
	Total	51	2752	2803

•Percent Correct = $(0+2752)/2803 = 98.2\%$

2 x 2 Contingency Table

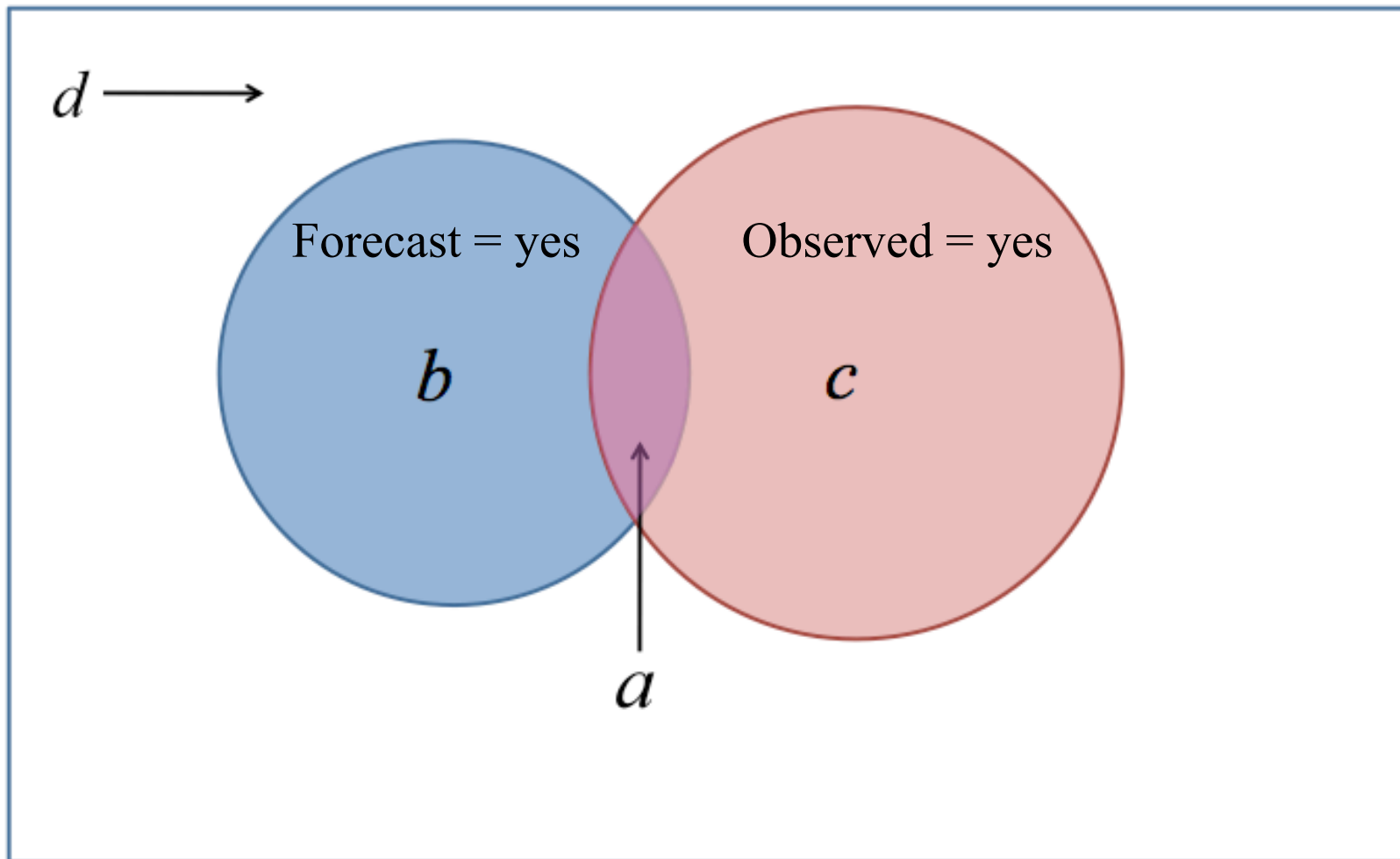
		Observed		
		Yes	No	Total
Forecast	Yes	Hit	False Alarm	Forecast Yes
	No	Miss	Correct Negative	Forecast No
	Total	Obs. Yes	Obs. No	Total

Common – though not universal - notation

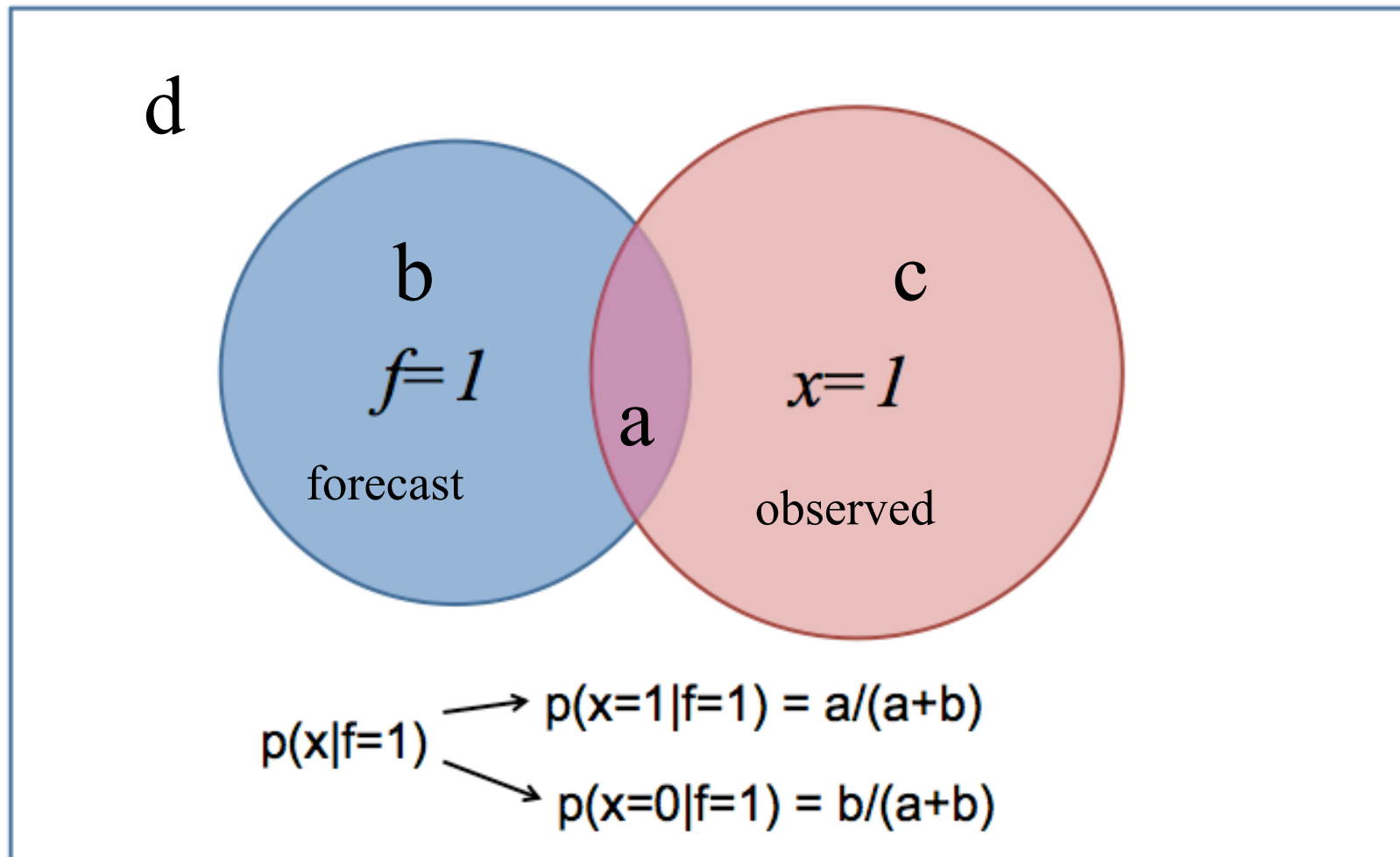
		Observed		
		Yes	No	Total
Forecast		a	b	a+b
	Yes	c	d	c+d
	No	a+c	b+d	n
	Total			

Base Rate (aka sample climatology) =
 $(a+c)/n$

Alternative Perspective on Contingency Table



Conditioning on Forecast

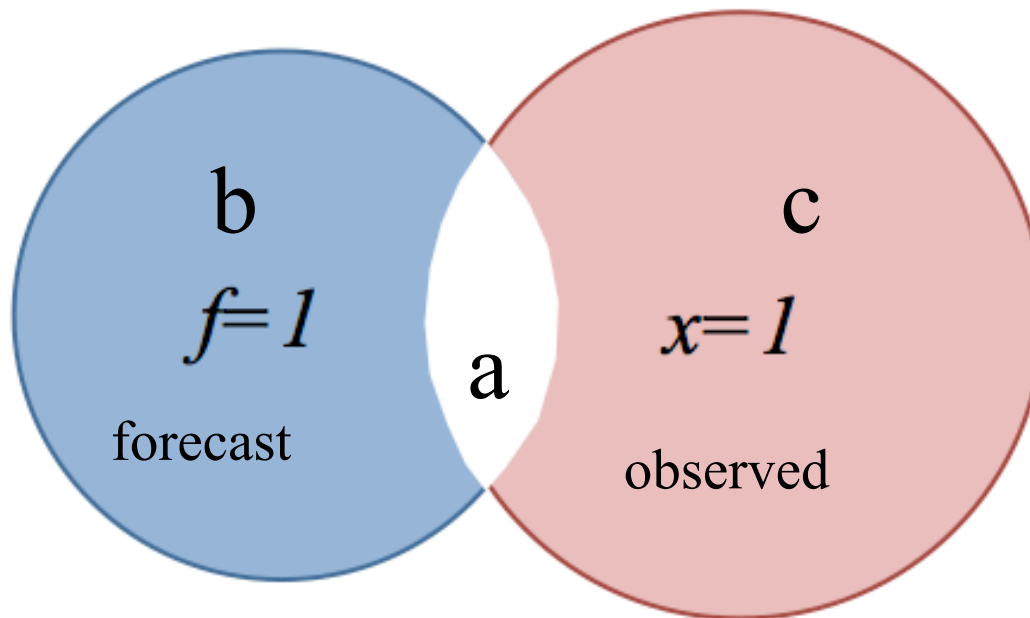


“good” forecasts

- $p(x=1 | f=1) = a/(a+b)$ to be as large as possible
 - fraction of “hits” in forecast region
- $p(x=0 | f=1) = b/(a+b)$ to be as small as possible
 - fraction of “false alarms” in forecast region

Conditioning on outcome

d



$$p(f|x=1) \begin{cases} \rightarrow p(f=1|x=1) = a/(a+c) \\ \rightarrow p(f=0|x=1) = c/(a+c) \end{cases}$$

“good” forecasts

- $p(f=1 | x=1)$ to be as large as possible
 - fraction of “hits” in observed region
- $p(f=0 | x=1)$ to be as small as possible
 - fraction of “missed” in observed region
- provides information regarding the ability of the forecast system to “discriminate” observed events vs. non-events
- these conditional probabilities are also called the “likelihoods” associated with the forecast

A laundry list of scores – most based on conditioning

- Hit Rate = $a/(a+c)$
- False Alarm Rate (POFD) = $b/(b+d)$
- False Alarm **Ratio** (FAR) = $b/(a+b)$
- Bias = $(a+b)/(a+c)$
- Threat Score or Critical Success Index = $a/(a+b+c)$
- PODn = $d/(b+d) = (1 - \text{POFD})$

		Observed		
		Yes	No	Total
Forecast	Yes	a	b	a+b
	No	c	d	c+d
	Total	a+c	b+d	n

Alternative Statistics

		Observed		
Forecast		Yes	No	Total
	Yes	28	72	100
	No	23	2680	2703
	Total	51	2752	2803

$$\text{Threat Score} = 28 / (28 + 72 + 23) = 0.228 \text{ (0)}$$

$$\text{Probability of Detection} = 28 / (28 + 23) = 0.55 \text{ (0)}$$

$$\text{False Alarm Ratio} = 72 / (28 + 72) = 0.720 \text{ (1)}$$

Skill Scores

(How do you compare the skill of easy to predict events with difficult to predict events?)

- Single value to summarize performance.
- Reference forecast - best naive guess; persistence, climatology.
- Proper skill scores – reflect forecaster true intent.
- A perfect forecast implies that the object can be perfectly observed.
- Reference forecast must be comparable.

Generic Skill Score

$$SS = \frac{(A - A_{ref})}{(A_{perf} - A_{ref})}$$

$$MSESS = 1 - \frac{MSE}{MSE_{limo}}$$

- Positively oriented – Positive is good

Calculation of Empirical ROC

Does not need to be a probability!

Does not need to be calibrated!

PROB	# YES	# NO
0.05	6	32
0.15	7	8
0.25	2	8
0.35	7	9
0.45	4	9
0.55	15	5
0.65	10	10
0.75	12	3
0.85	16	8
0.95	146	14

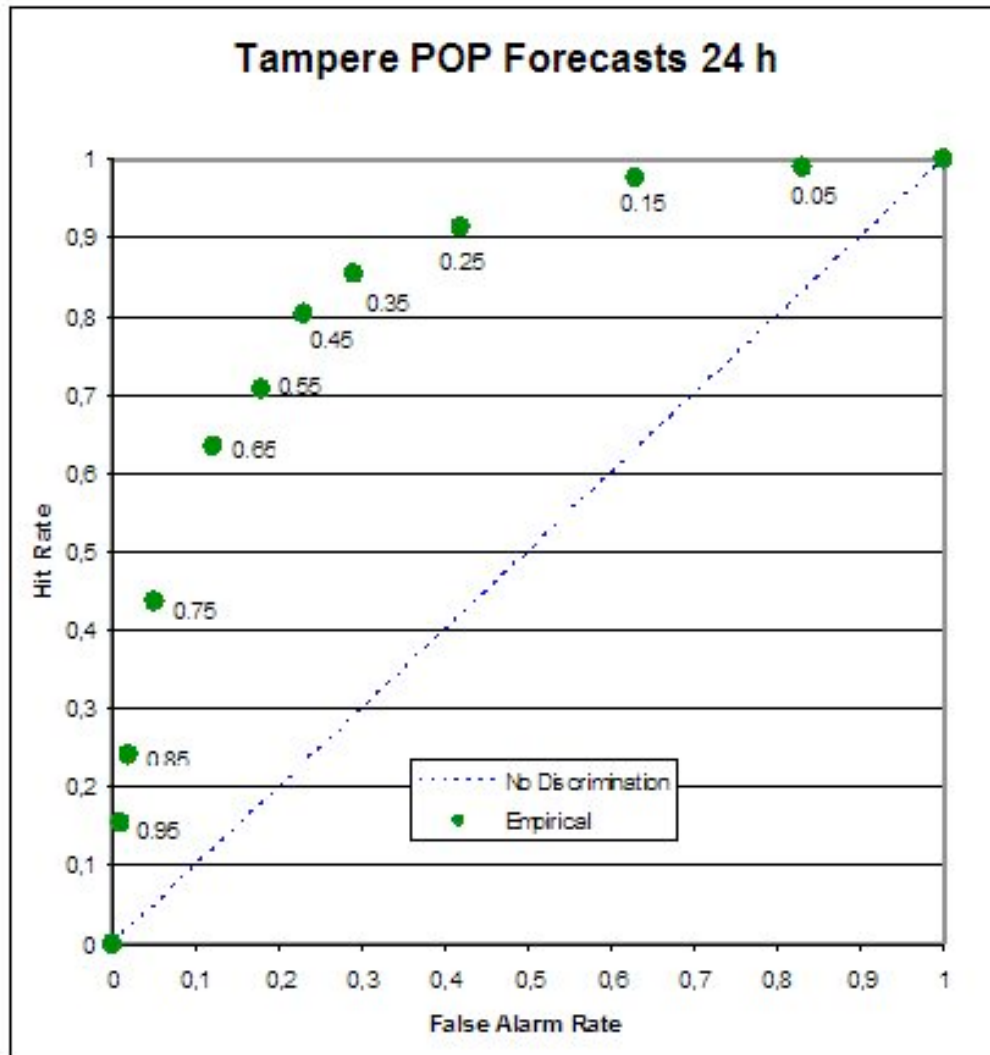
c

d

a

b

Empirical ROC

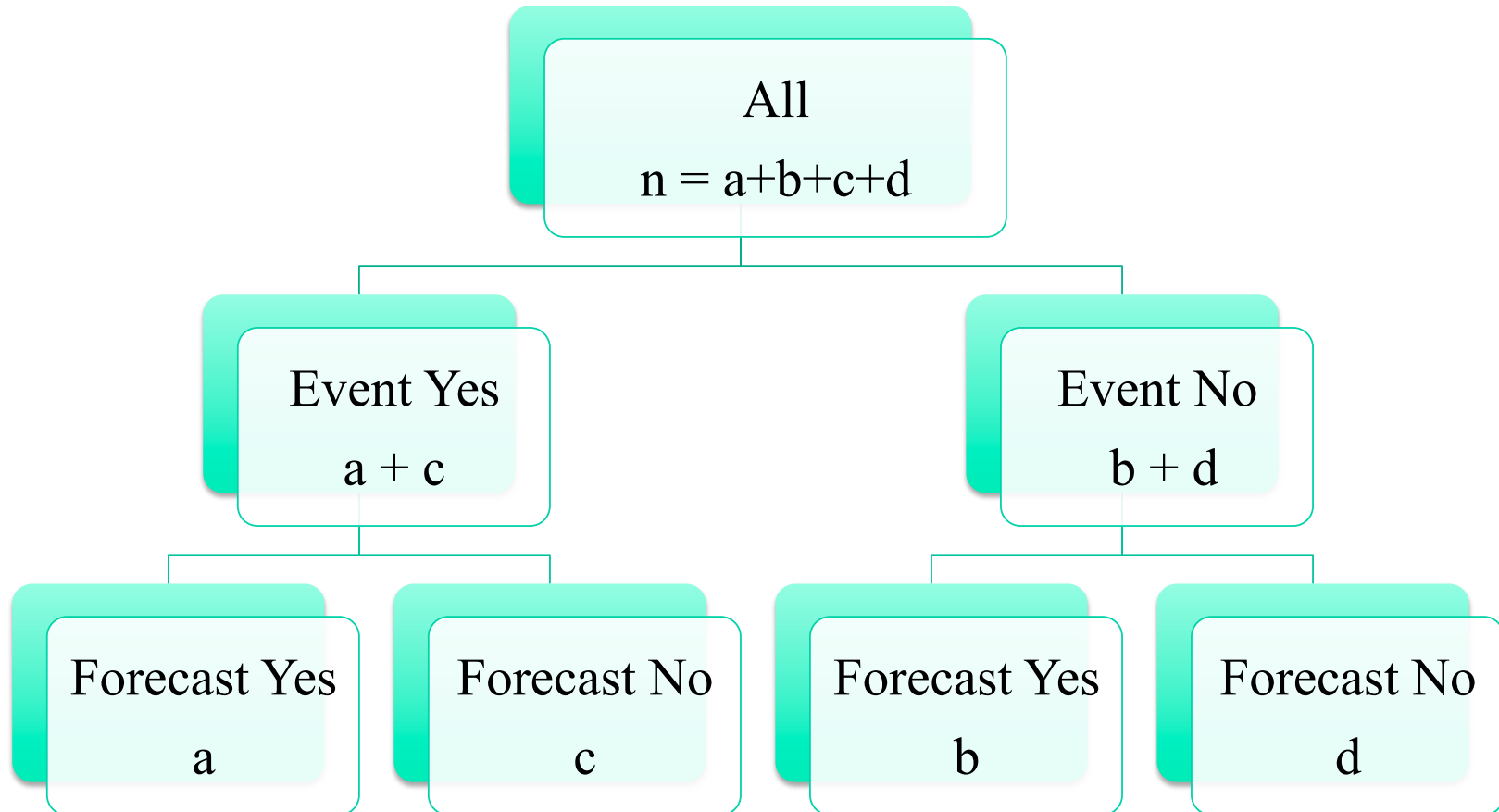


- Area under the curve is a useful measure; Perfect = 1, random = 0.5
- May fit a distribution to the curve
- Forecasts needn't be calibrated
- Interpretation conditioned on base rate

Some cautions about conditional probabilities.

- People don't understand them!
- There are always two ways to present data from a contingency table - as conditional probabilities and as a natural frequency.

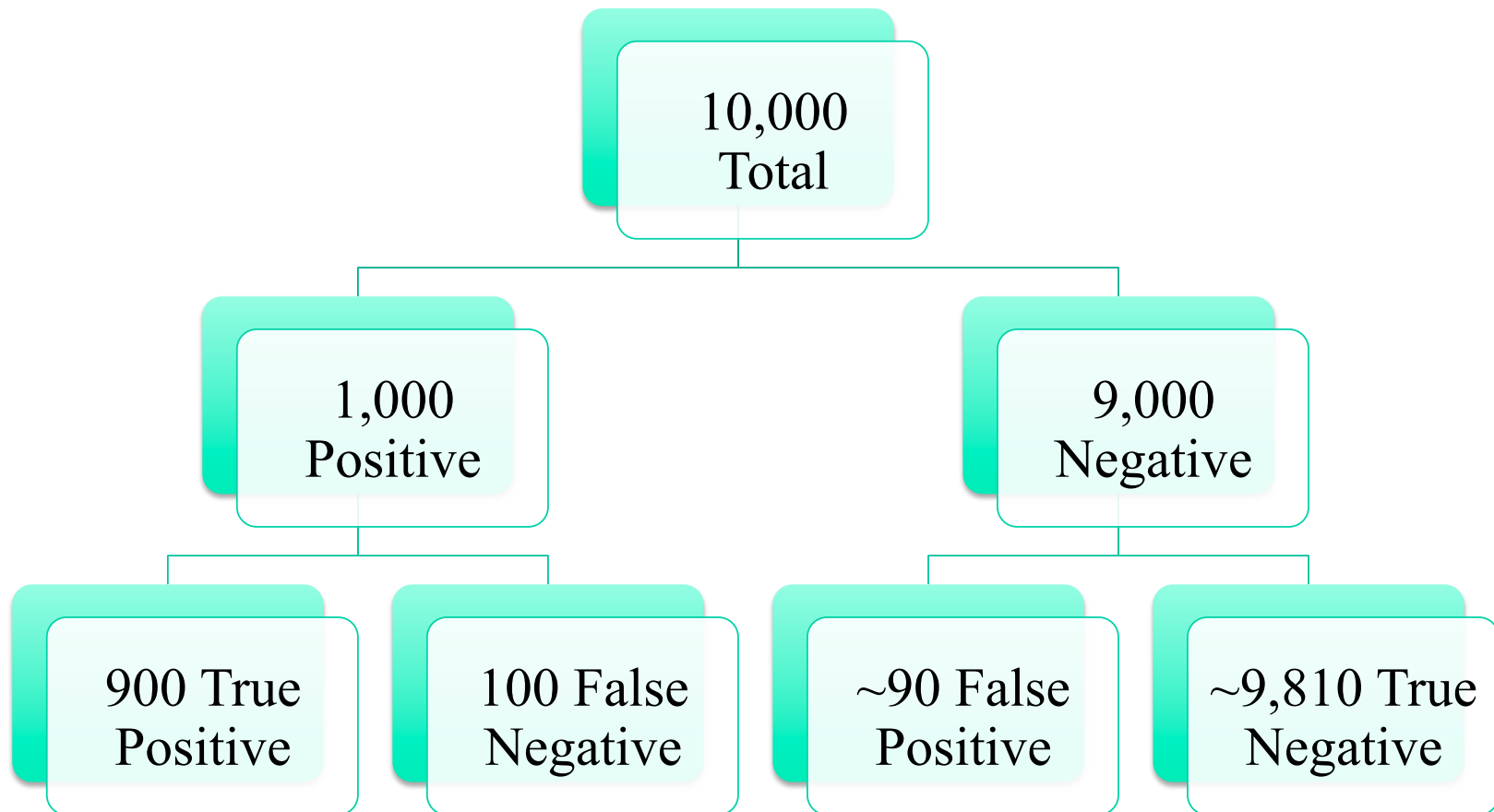
Still another perspective



10,000 Observations

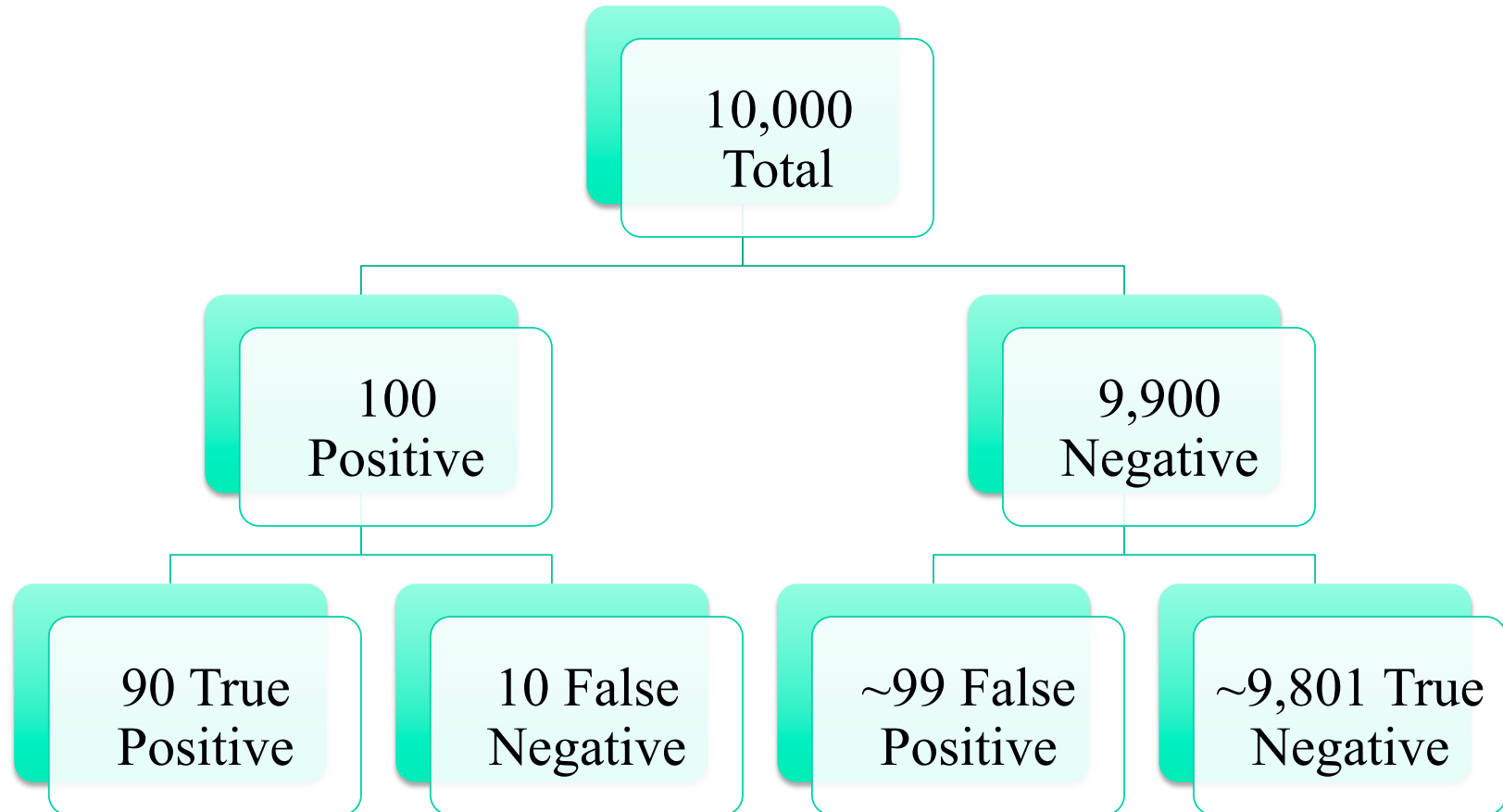
(Base Rate 10% = 1,000 pos/ 10,1000 total)

$POD_y = 0.9$, $POD_n = 0.99$



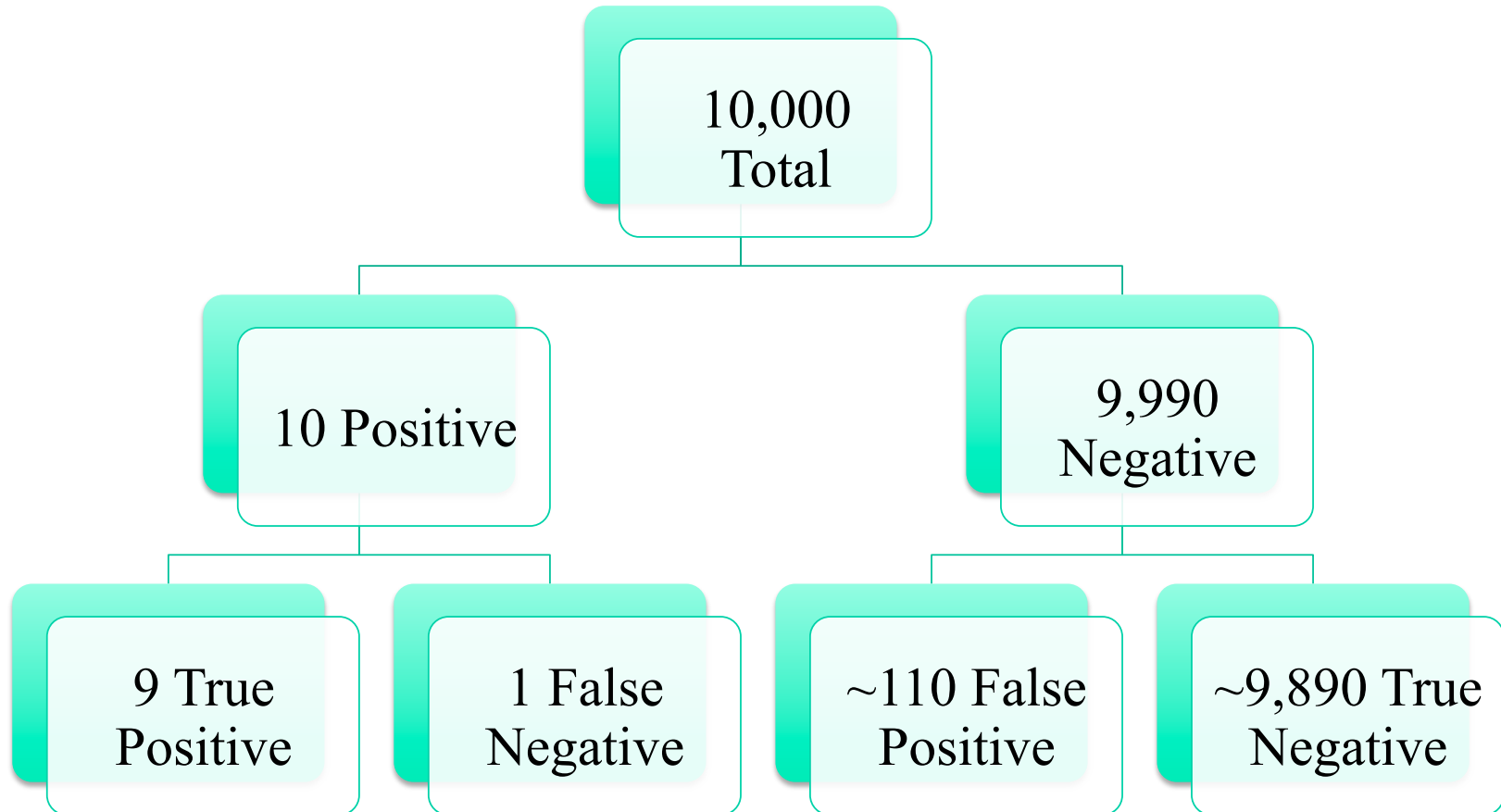
10,000 Observations (Base Rate 1%)

$$POD_y = 0.9, \text{ } POD_n = 0.99$$



10,000 Observations (Base Rate 0.1%)

$$POD_y = 0.9, POD_n = 0.99$$



What is verification bias?

- To correctly calculate POD_y and POD_n , one assumes the prevalence of turbulence among the observed population is equivalent to the unobserved.
- There is no external information about turbulence would make it more likely to be observed. This is known as the missing at random (MAR) assumption.

Often ignored, but important assumptions

- Data is independent
- Sample data reflects the population data
- Data collection is unbiased
 - Do sites reflect the larger domain?
 - Does a human selectively report values
 - Does sampling frequency catch temporal changes at the same resolution?

Example of verification bias

Likely to be reported

	Observed		
Forecast		Yes	No
Yes		40	10
No		5	0

Less likely to be reported

	Observed		
Forecast		Yes	No
Yes		0	85
No		5	855

All data reported

	Observed		
Forecast		Yes	No
Yes		40	95
No		10	855

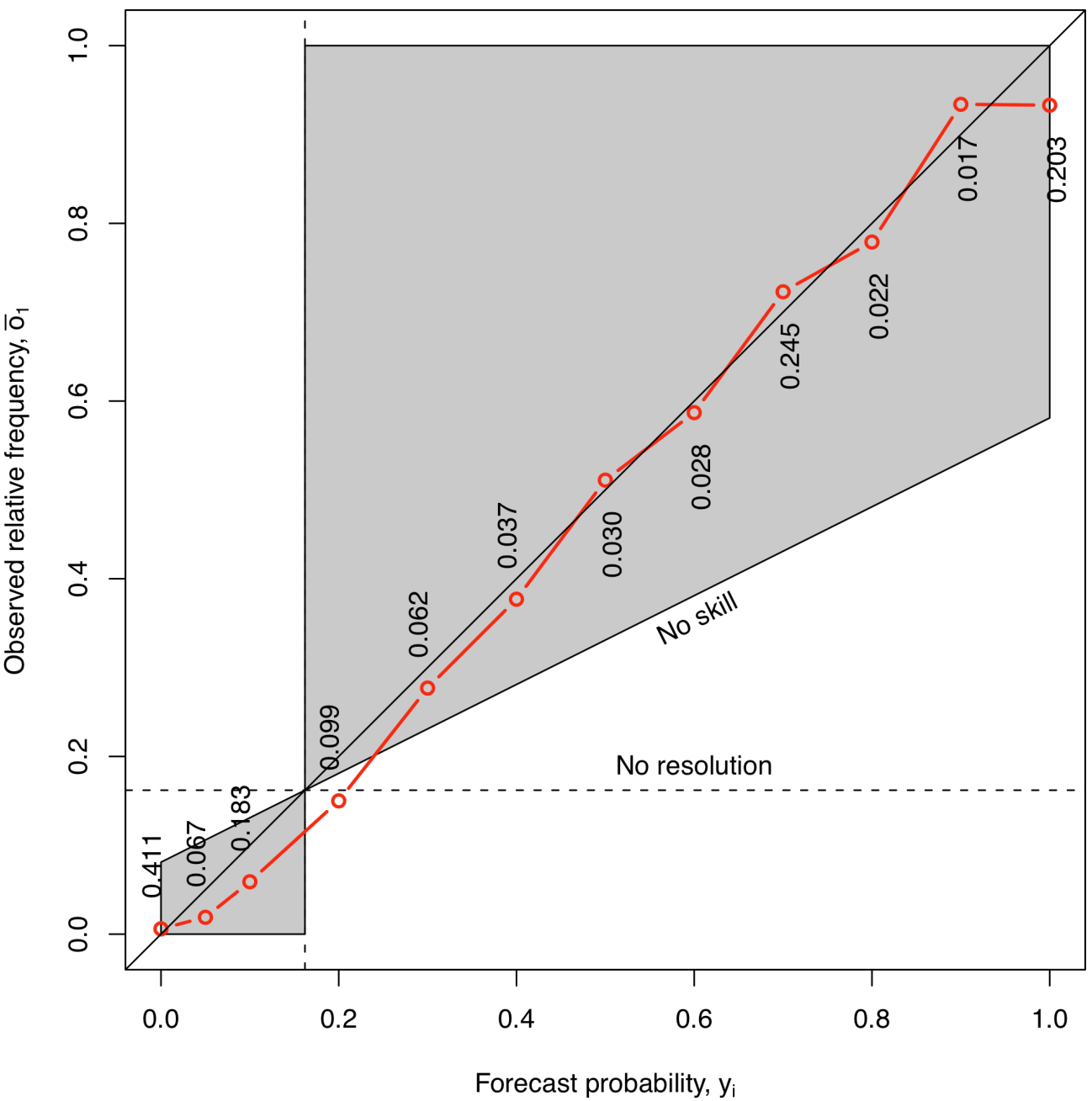
Some data reported

	Observed		
Forecast		Yes	No
Yes		0	85
No		2	273

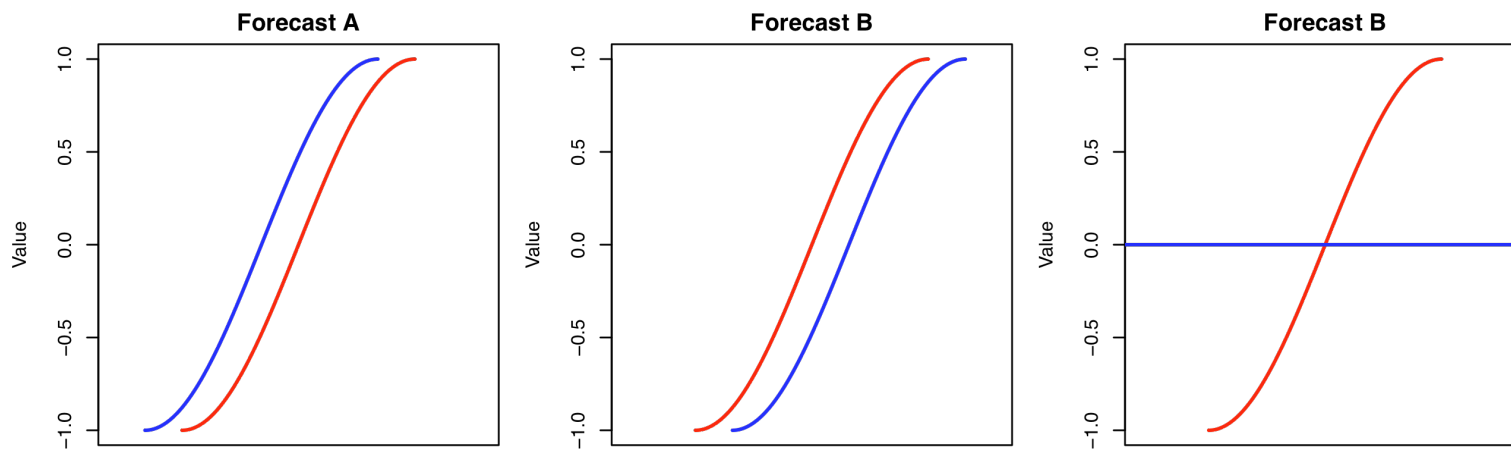
	All Data	Combined stratified data
POD _y	0.8	0.85
False positive	0.1	0.26

Attribute Diagram

More tomorrow



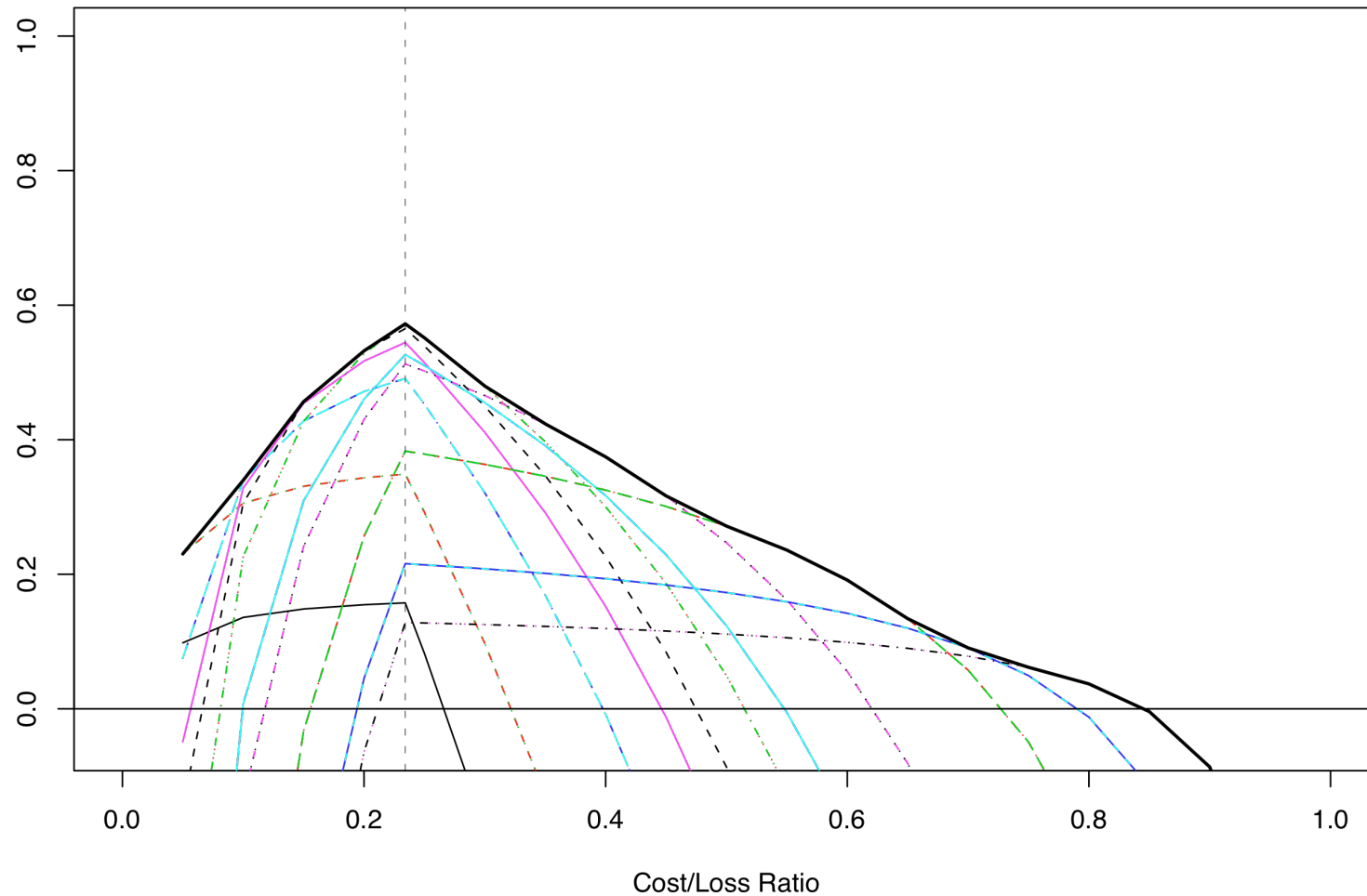
User specific verification metrics



In the context of forecasting ramps, timing is crucial and is not captured by MAE.

Value based metrics

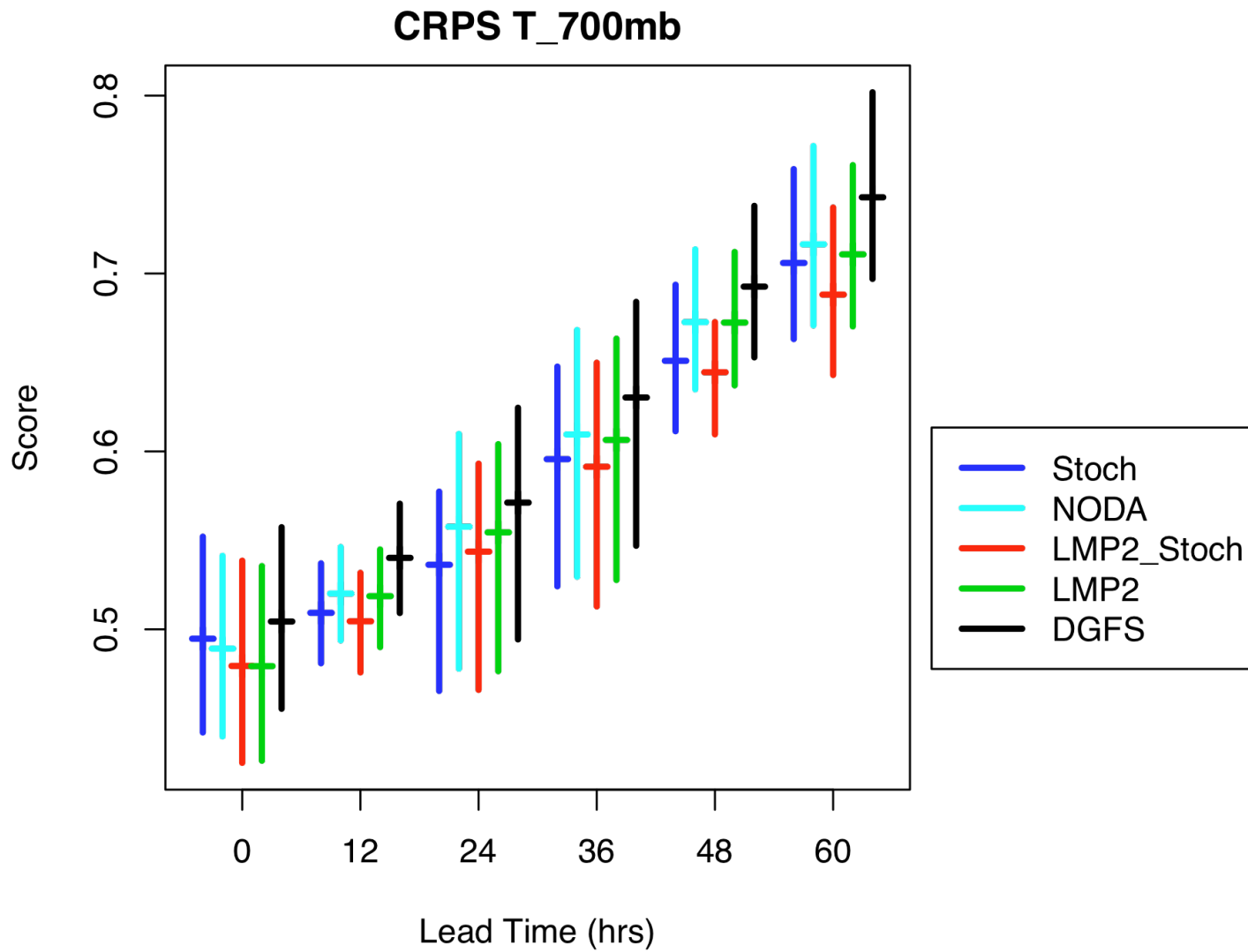
- All forecasts with skill do not have value!



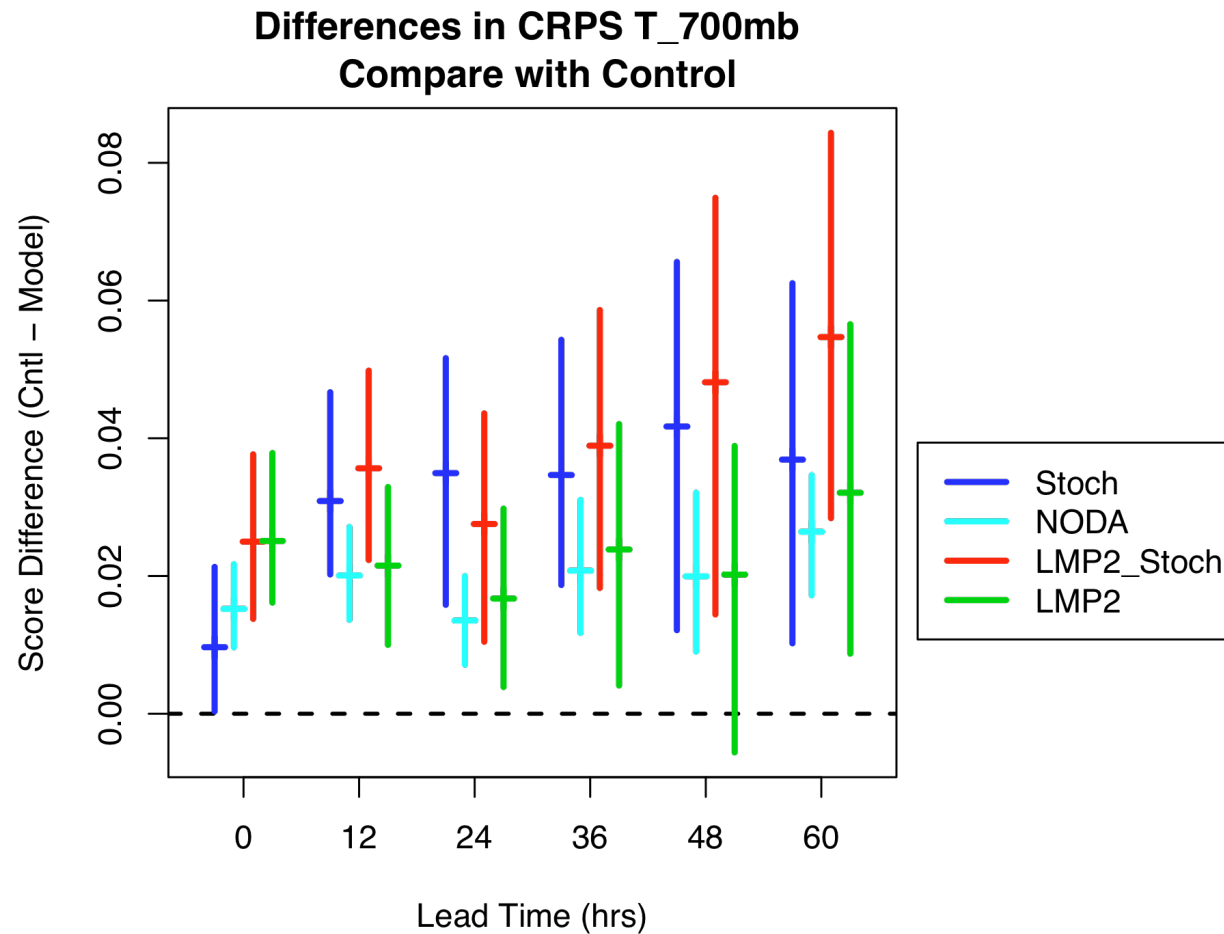
Sampling variability can be used

- When you are comparing two forecasts of the same event, evaluate the differences.
- Sampling variability is large and can quickly overwhelm small, but significant differences.

Scores in absolute terms



Examining the distribution of differences



Verification software

- MET provides data necessary to calculate verification statistics
- Most verification statistics discussed here rely on basic aggregate type commands found in most statistical programming languages such as R, Splus, Matlab, IDL, ...

R and Verification Package

- Open source, runs on any operating system.
- Complete set of basic statistical functions similar to SAS, Matlab, ...
- > 2,500 contributed packages including verification.
 - All basic verification functions outlined in Wilks, plus contributed code.

references

- Jolliffe and Stehenson (2003): Forecast verification: A practitioner's guide, Wiley & sons
- Nurmi (2003): Recommendations on the verification of local weather forecasts. ECMWF Technical Memorandum, no. 430
- Wilks (2006): Statistical methods in the atmospheric sciences, ch. 7. Academic Press
- JWGFVR (2009): Recommendation on verification of precipitation forecasts. WMO/TD report, no.1485 WWRP 2009-1
- <http://tinyurl.com/verif-training>
- http://www.bom.gov.au/bmrc/wefor/staff/eee/verif/verif_web_page.html