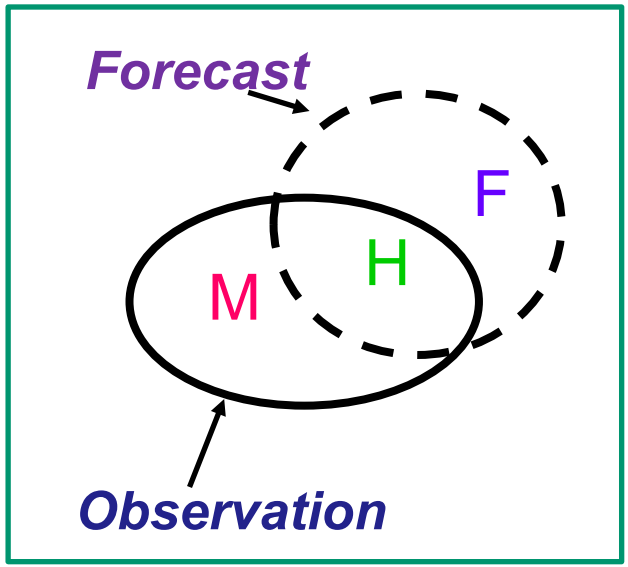
GSS

CSI

Freq Bias



Categorical Verification

Tina Kalb

Contributions from Tara Jensen, Matt Pocernich, Eric Gilleland,
Tressa Fowler, Barbara Brown and others

Base Rate

Hit Rate

FAR

POD Y

Finley Tornado Data (1884)

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

John P. Finley.

Forecast answering the question:

Will there be a tornado?

YES

NO

Observation answering the question:

Did a tornado occur?

YES

NO

Answers fall into 1 of 2 categories ⁂ Forecasts and Obs are Binary

Finley Tornado Data (1884)

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

John P. Finley.

		Observed		
Forecast		Yes	No	Total
	Yes			
	No			
	Total			

Contingency Table

A Success?

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

John P. Finley.

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

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

What if forecaster never forecasted a tornado?

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

John P. Finley.

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

Percent Correct = (0+2752)/2803 = 98.2% !!!!

maybe Accuracy is not the most informative statistic

But the contingency table concept is good…

2 x 2 Contingency Table

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

Example: Accuracy = (**Hits**+**Correct Negs**)/**Total**

MET supports both 2x2 and NxN Contingency Tables

Common Notation

(however not universal notation)

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

Example: Accuracy = (a+d)/n

What if data are not binary?

Threshold

Examples:
Temperature < 0 C
Precipitation > 1 inch
CAPE > 1000 J/kg
Ozone > 20 µg/m³
Winds at 80 m > 24 m/s
500 mb HGTS < 5520 m
Radar Reflectivity > 40 dBZ
MSLP < 990 hPa
LCL < 1000 ft
Cloud Droplet Concentration > 500/cc

Hint: Pick a threshold that is meaningful to your end-user

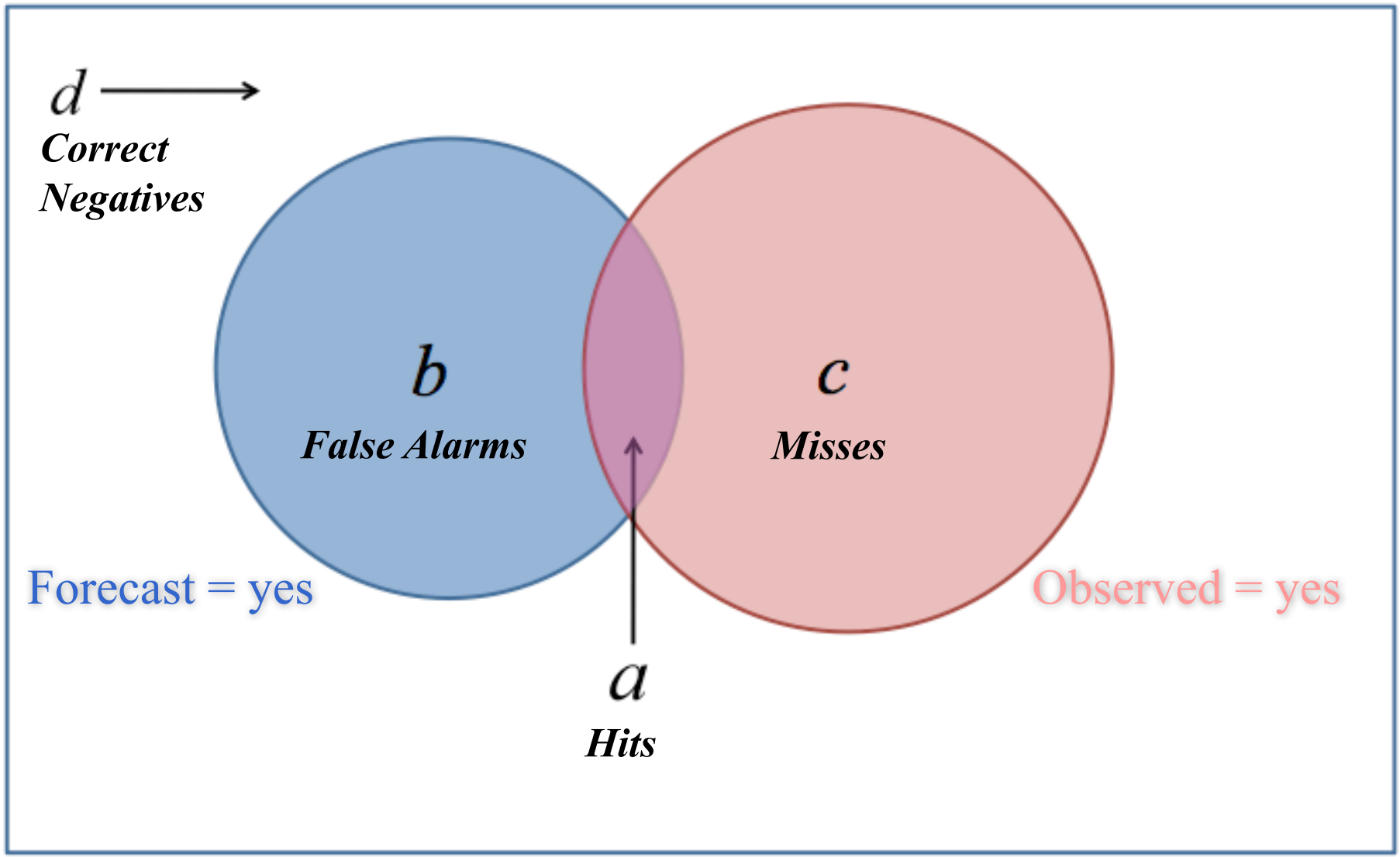
Contingency Table for Freezing Temps (i.e. T<=0 C)

		Observed		
Forecast		<= 0C	> 0C	Total
	<= 0C	a	b	a+b
	> 0C	c	d	c+d
	Total	a+c	b+d	n

Another Example:

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

Alternative Perspective on Contingency Table



Conditioning to form a statistic

- Considers the probability of one event given another event
- Notation: p(X|Y=1) is probability of X occuring given Y=1 or in other words Y=yes

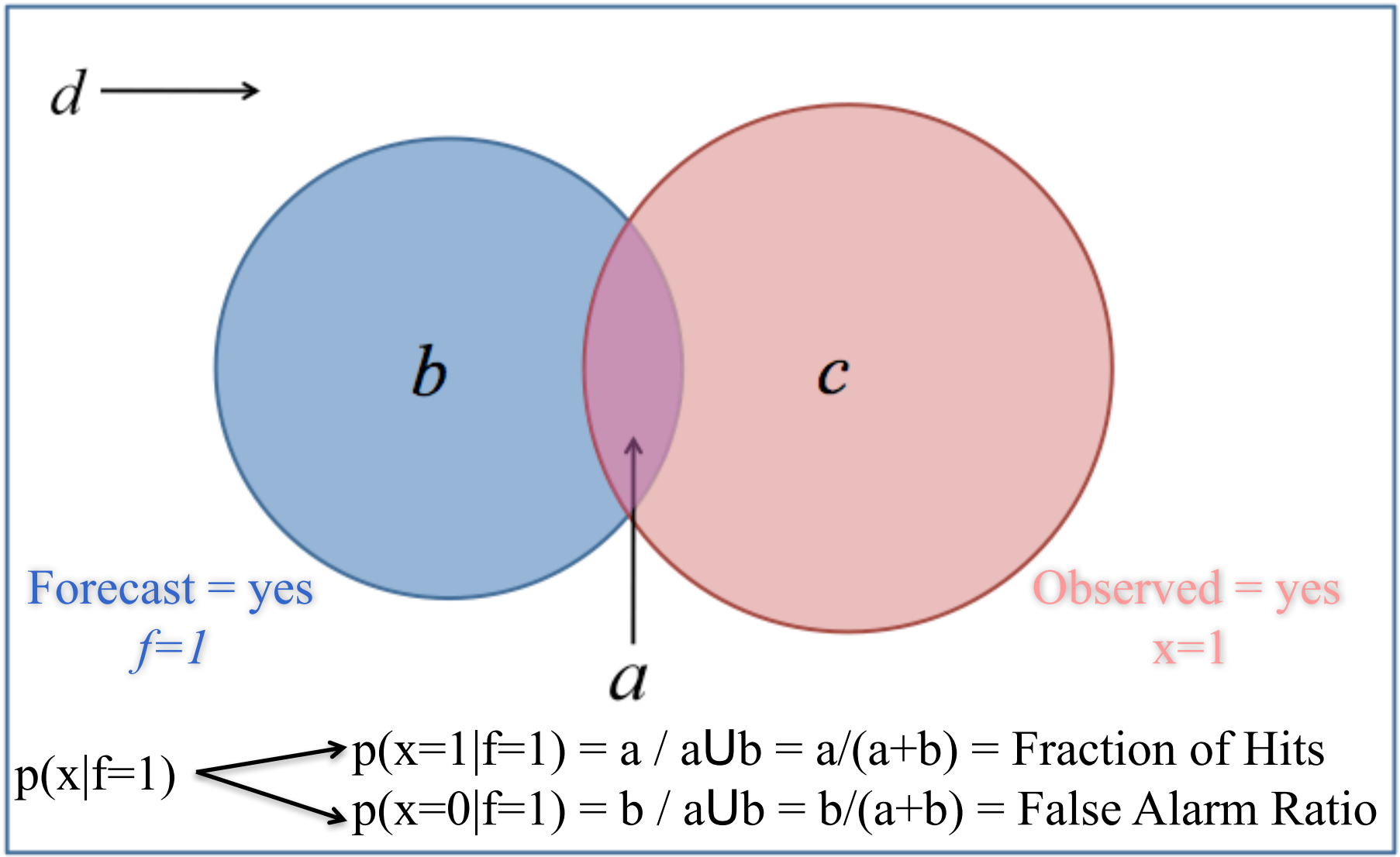
Conditioning on Fcst provides:

- Info about how your forecast is performing
- Apples-to-Oranges comparison if comparing stats from 2 models

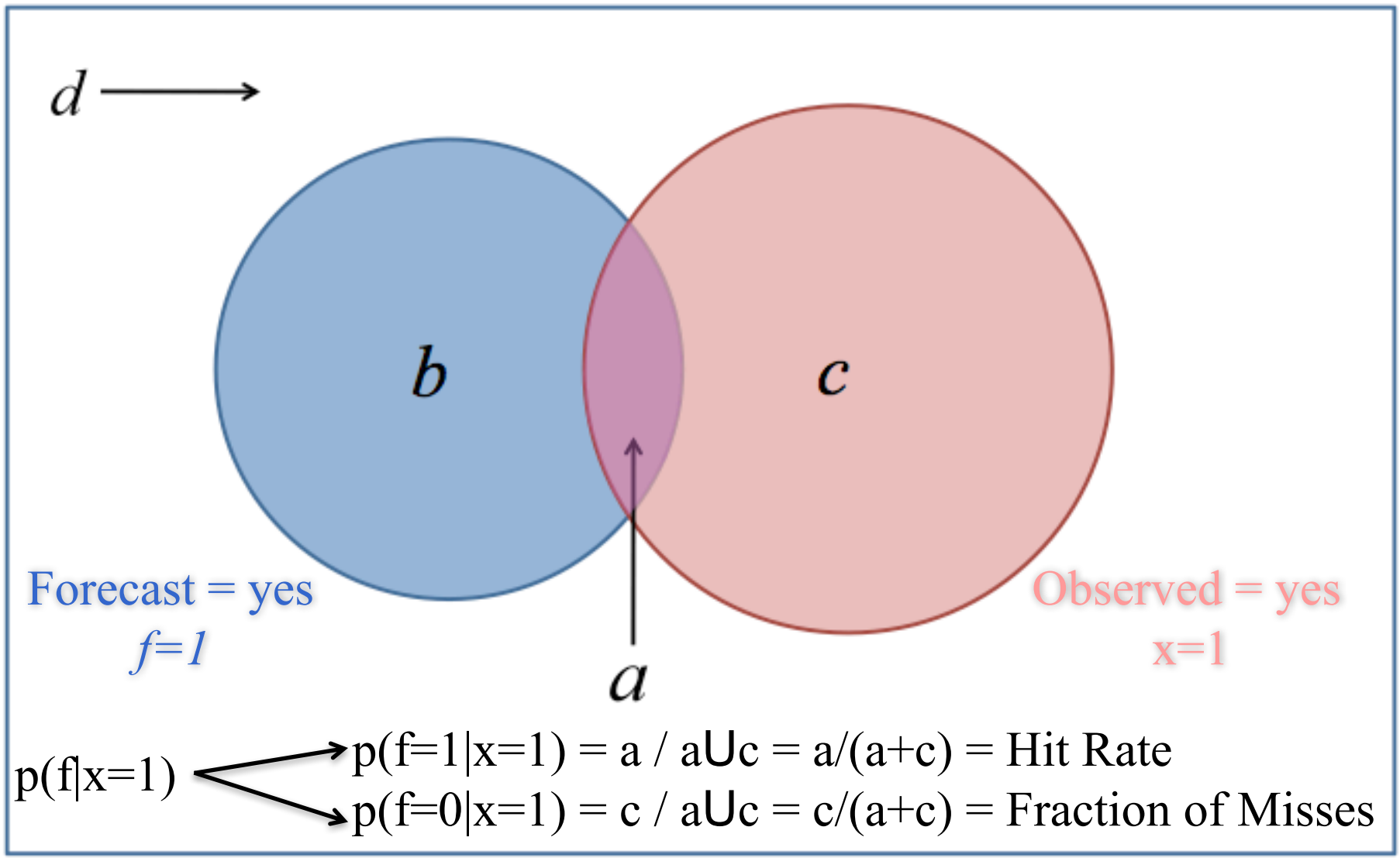
Conditioning on Obs provides:

- Info about ability of forecast to discriminate between event and non-event - also called Conditional Probability or "Likelihood"
- Apples-to-Apples comparison if comparing stats from 2 models

Conditioning on forecasts



Conditioning on observations



What's considered good?

Conditioning on Forecast

Fraction of hits - $p(x=1|f=1) = a/(a+b)$: close to 1

False Alarm Ratio - $p(x=0|f=1) = b/(a+b)$: close to 0

Conditioning on Observations

Hit Rate - $p(f=1|x=1) = a/(a+c)$: close to 1

[aka Probability of Detection Yes (PODy)]

Fraction of misses $p(f=0|x=1) = a/(a+c)$: close to 0

Examples of Categorical Scores

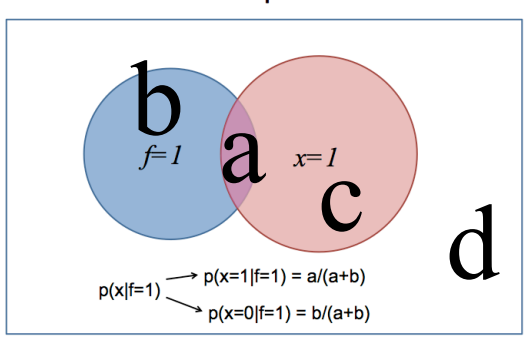
(most based on conditioning)

- PODy = a/(a+c)
- False Alarm **Ratio** (FAR) = b/(a+b)
- PODn = d/(b+d) = (1 – POFD)
- False Alarm Rate (POFD) = b/(b+d)
- (Frequency) Bias (FBIAS) = (a+b)/(a+c)
- Threat Score or Critical Success Index = a/(a+b+c) (CSI)

POD
Probability of Detection

POFD
Probability of False Detection

Conditional probabilities



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

Examples of CTC calculations

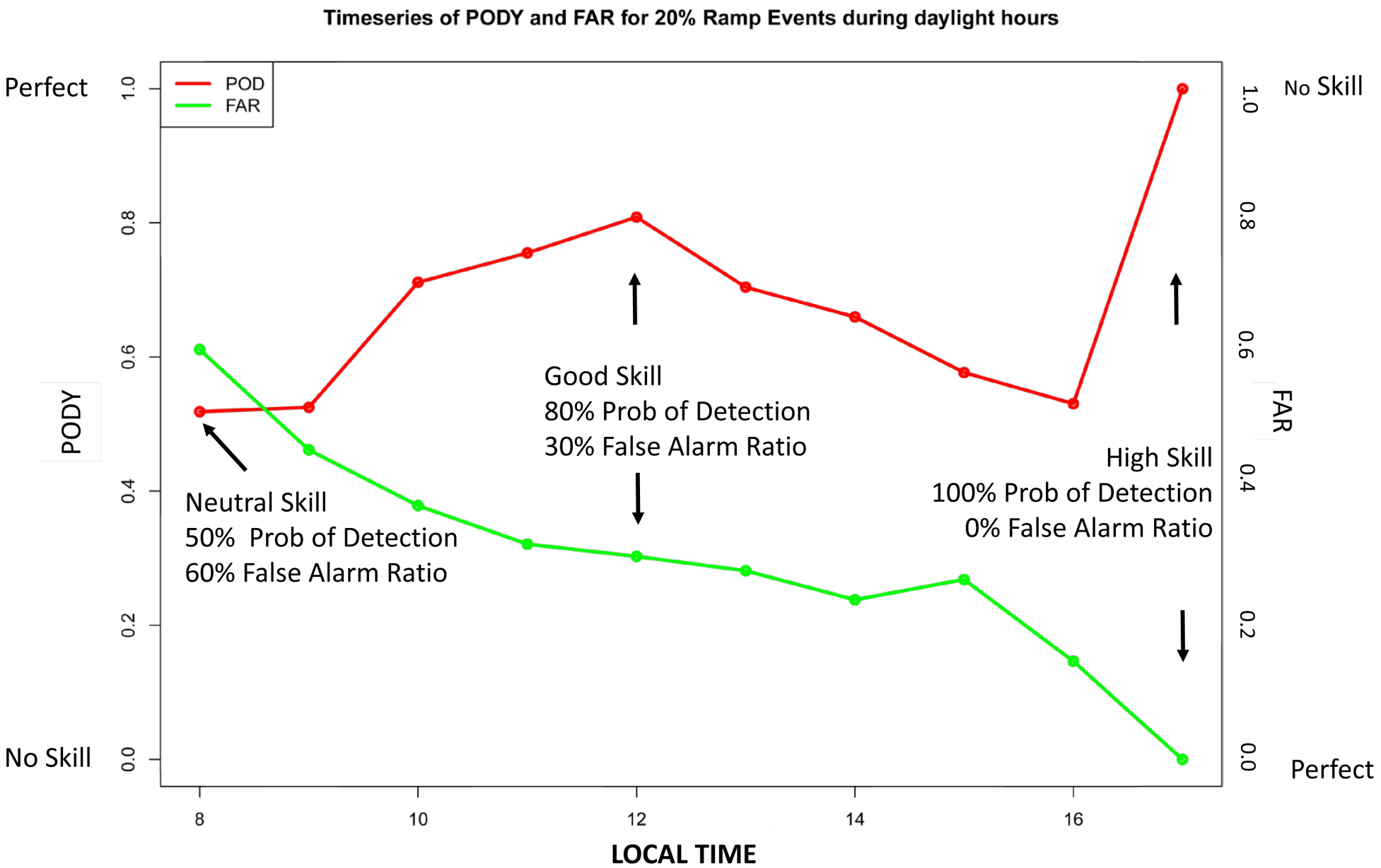
		Observed		
Forecast		Yes	No	Total
	Yes	28	72	100
	No	23	2680	2703
	Total	51	2752	2803

Threat Score = 28 / (28 + 72+ 23) = 0.228

Probability of Detection = 28 / (28 + 23) = 0.55

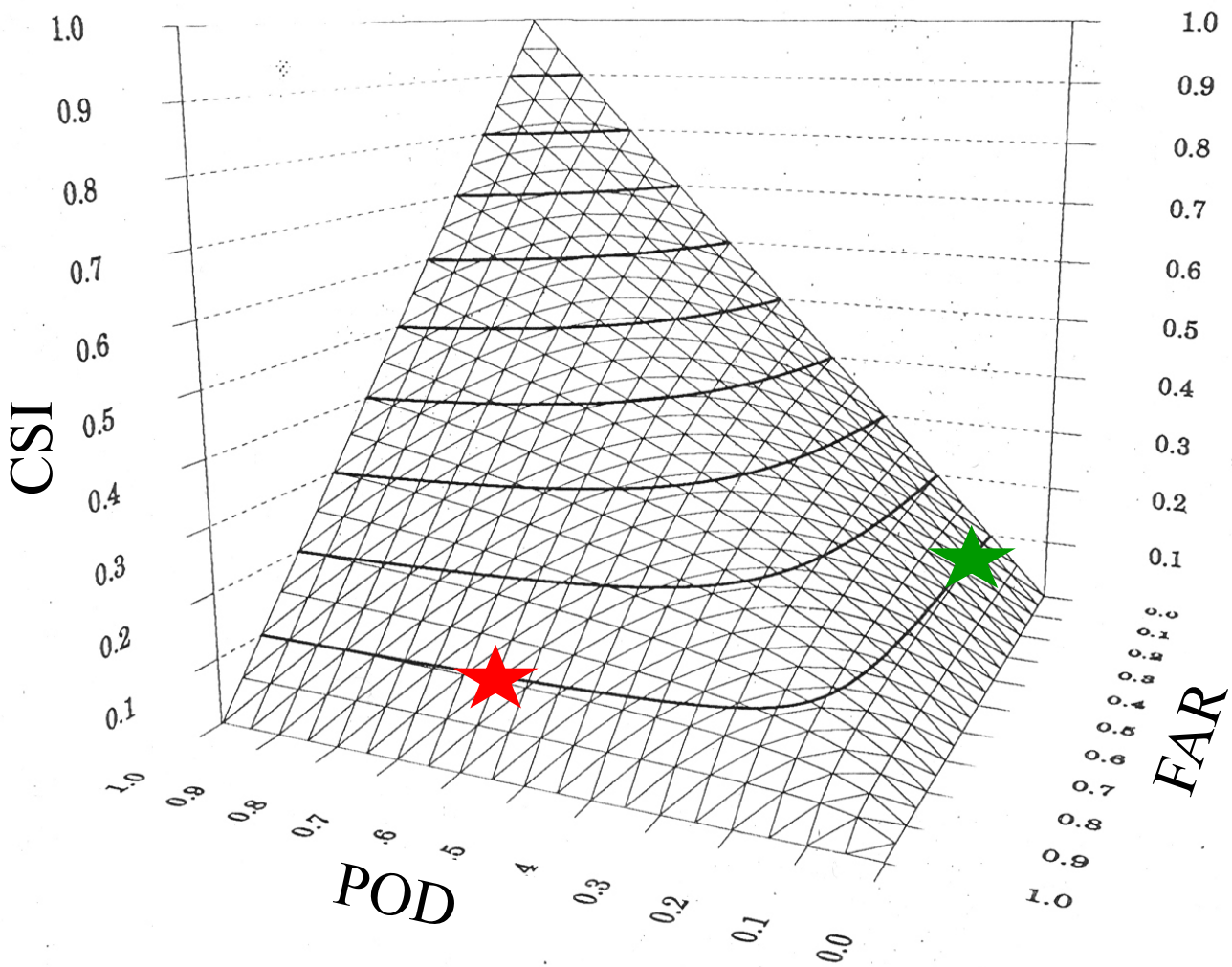
False Alarm Ratio= 72/(28 + 72) = 0.720

Example



Relationships among scores

- CSI is a *nonlinear* function of POD and FAR
- CSI depends on base rate (event frequency) and Bias



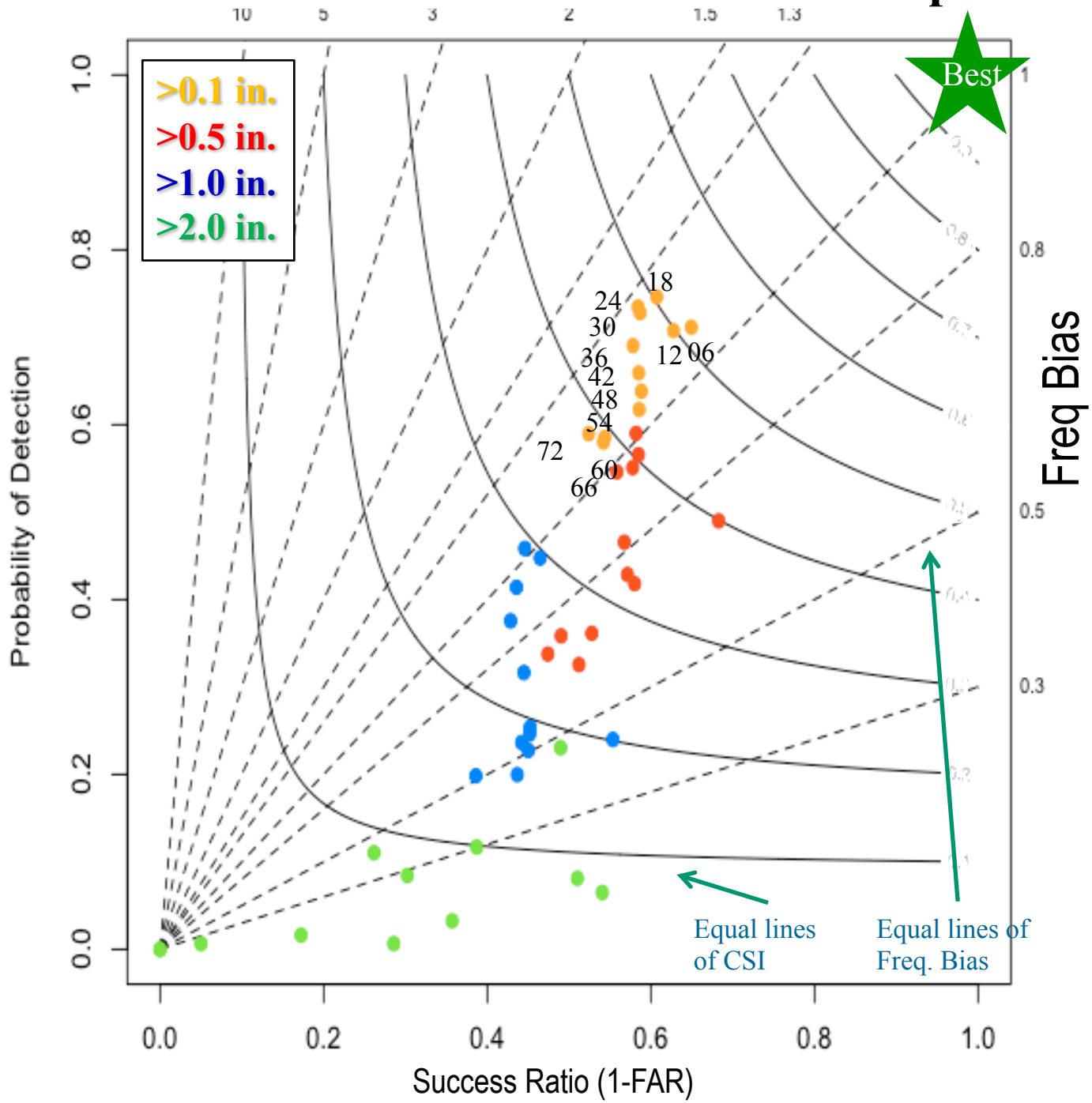
$$\mathrm{CSI} = \frac{1}{\frac{1}{\mathrm{POD}} + \frac{1}{1-\mathrm{FAR}} - 1}$$

$$\mathrm{Bias} = \frac{\mathrm{POD}}{1-\mathrm{FAR}}$$

Very different combinations of FAR and POD can lead to same CSI value

HMT Performance Diagram

- On same plot
  - POD
  - 1-FAR (aka Success Ratio)
  - CSI
  - Freq Bias
- Dots: Scores Aggregated Over Lead Time
- Colors: different thresholds
- Results:
  - Decreasing skill with higher thresholds across multiple metrics
  - Highest skill 18 – 24 h lead times



Roberts et al. (2011), Roebber (WAF, 2009), Wilson (presentation, 2008)

Skill Scores

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

- Provides a single value to summarize performance.
- Reference forecast - best naive guess; persistence; climatology.
- Reference forecast must be comparable.
- Perfect forecast implies that the object can be perfectly observed.

Generic Skill Score

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

where A = any measure
ref = reference
perf = perfect

Example: $$MSESS = 1 - \frac{MSE}{MSE_{climo}}$$

where MSE = Mean Square Error

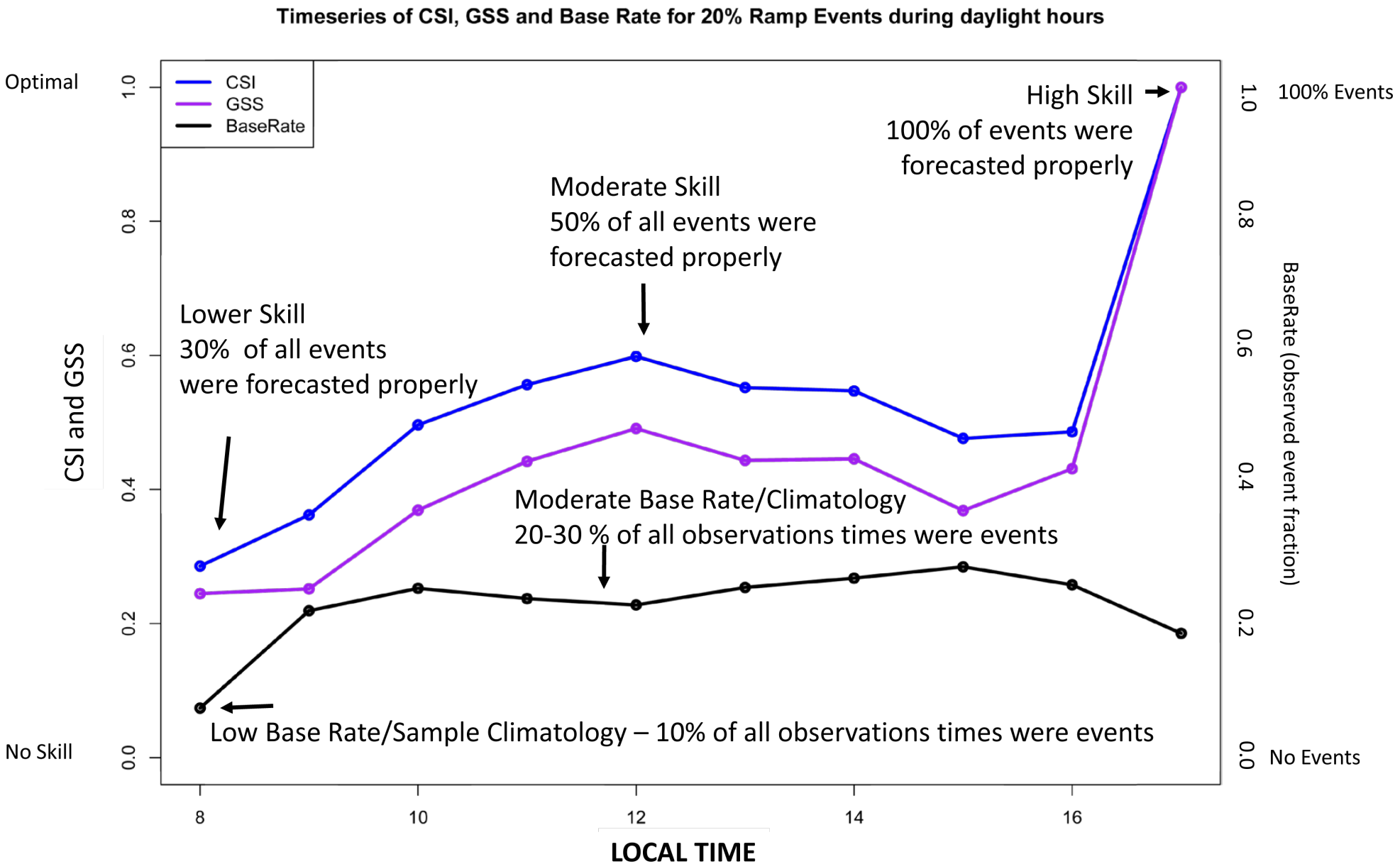
- **Interpreted as fractional improvement over reference forecast**
  - Reference could be: Climatology, Persistence, your baseline forecast, etc..
  - Climatology could be a separate forecast or a gridded forecast sample climatology
- **SS typically positively oriented with 1 as optimal**

Commonly Used Skill Scores

- **Gilbert Skill Score -** based on the CSI corrected for the *number of hits expected by chance*.
- **Heidke Skill Score -** based on Accuracy corrected by the *number of hits expected by chance*.
- **Hanssen-Kuipers Discriminant** – (Pierce Skill Score) measures ability of forecast to discriminate between (or correctly classify) events and non-events. H-K=POD-POFD

- **Brier Skill Score** for probabilistic forecasts
- **Fractional Skill Score** for neighborhood methods
- **Intensity-Scale Skill Score** for wavelet methods

Example



Thank you!

References:

Jolliffe and Stephenson (2012): Forecast Verification: a practitioner's guide, Wiley & Sons, 240 pp.

Wilks (2011): Statistical Methods in Atmospheric Science, Academic press, 467 pp.

Stanski, Burrows, Wilson (1989) Survey of Common Verification Methods in Meteorology

http://www.eumetcal.org.uk/eumetcal/verification/www/english/courses/msgcrs/index.htm

WMO Verification working group forecast verification web page, http://www.cawcr.gov.au/projects/verification/