The Relationship Between ROC, Performance, and the Quality-Decision Threshold Diagrams

HAROLD BROOKS
NOAA/NATIONAL SEVERE STORMS LAB
HAROLD.BROOKS@NOAA.GOV
@HEBROOKS87
UNIV. OF OKLAHOMA SCHOOL OF METEOROLOGY

THANKS TO
JIMMY CORREIA AND BURKELY GALLO

How this got started

- Murphy (1993)-relationship between quality and value
- Early 2000s
  - Was a goal of POD=0.8, FAR=0.5 for tornado warnings reasonable?
  - “With our current science, there’s no excuse for an FAR>0.25”
- More recent
  - What happened with US tornado warning performance in 2012/3?

Basic premises

- Visualization of multiple aspects of forecast performance can help in understanding of system
  - Different diagrams emphasize/hide different things
  - Choices reflect implicit statement of values
    - "The numbers have no way of speaking for themselves. We speak for them. We imbue them with meaning."-N. Silver, *The Signal and the Noise*
  - I live in the world of rare events and short-term forecasts
- Use toy models of forecasting to understand relationships
- Comparison to "real" forecasts

Long-term goals

- Create a simple model of forecast systems that we can use to look at impacts of changes in any aspect
  - Improving science
  - Different user decision problems
  - Probabilistic forecasts that can be thresholded

Exploit 2x2 Tables

Quality

		Event	
		Y	N
Forecast	Y	a	b
	N	c	d

- POD=a/(a+c)
- POFD=b/(b+d)
- SR=1-FAR=b/(a+b)
- DFR=c/(c+d)
- Base rate=f=(a+c)/(a+b+c+d)

Value

		Event	
		Y	N
Action?	Y	A	B
	N	C	D

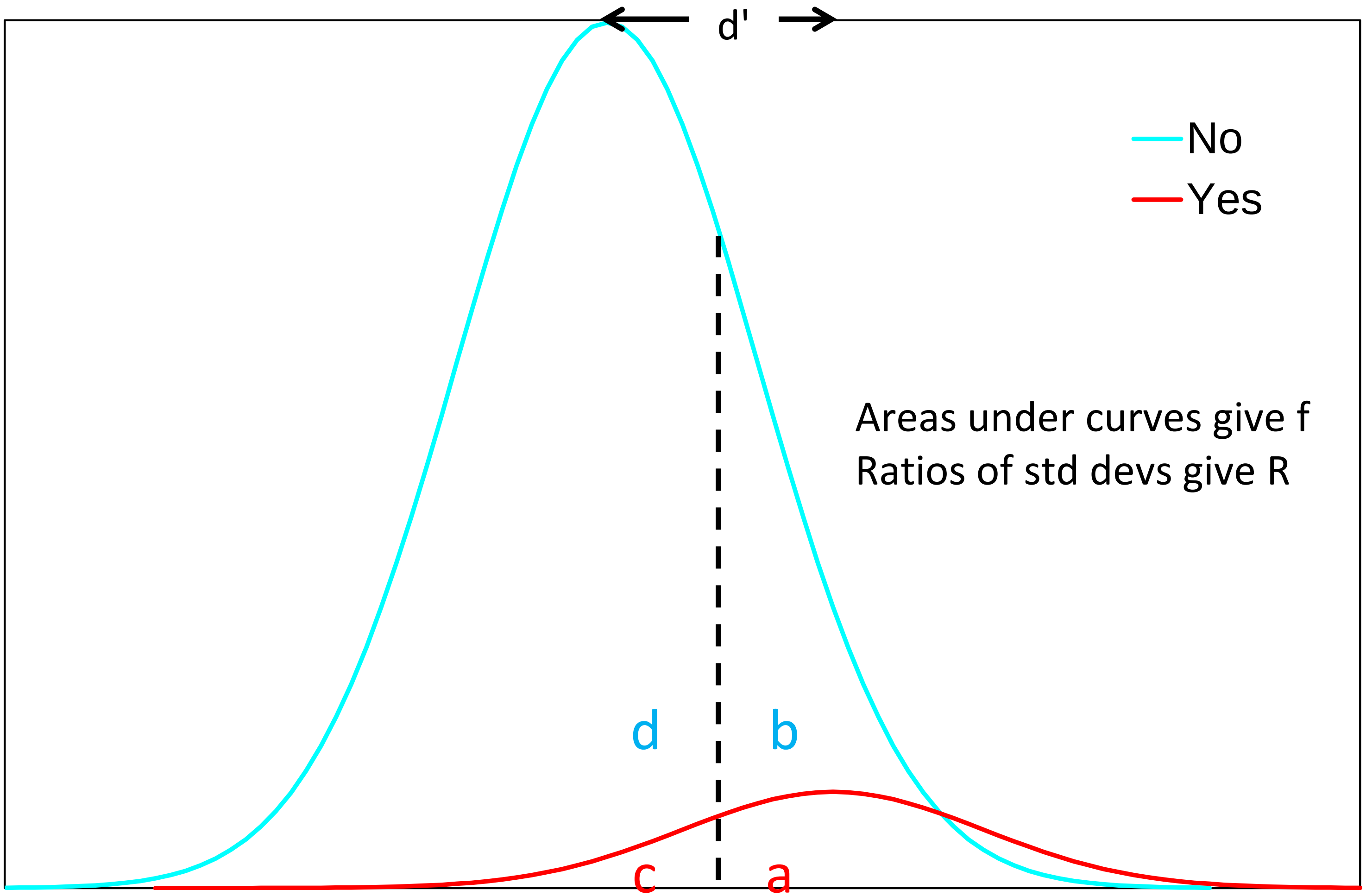
- Misclassification Cost Ratio (α)
- Act if $p > \alpha = \frac{(B-D)}{(B-D)+(C-A)}$
- Act if $p > \alpha = \frac{Cost(FA)}{Cost(FA)+Cost(ME)}$

Modelling the Problem

- Signal Detection Theory [following Mason (1982)]
  - Gaussian distributions for “yes” and “no” events, separated by d’
  - Ratio of standard deviations (R) $=\frac{\sigma_{no}}{\sigma_{yes}}$
  - Local separation (d*) comes from z(POD)-z(POFD)
    - If R=1, d*=d’ always

- f=base rate of event requiring decisions (needed to get all elements of table)

Modelling the Problem



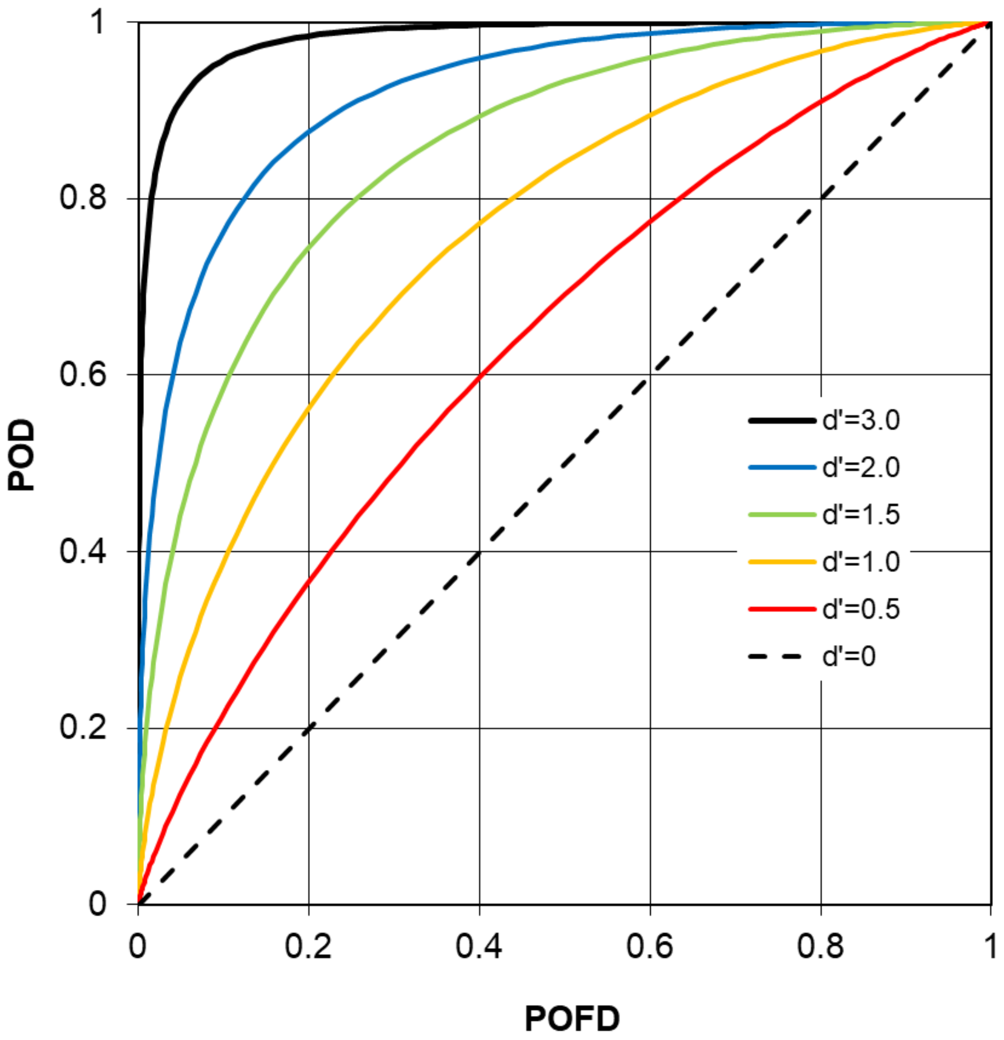
Basic diagrams today

- Relative operating characteristics (Mason 1982)
  - POD vs. POFD
  - No information on bias
  - For rare events, real forecasts typically cluster in low POFD
  - Also show z-transform diagram of POD and POFD
- Performance diagram (Roebber 2009)
  - Reversed axes from precision-recall curve
  - POD vs SR
  - No information on correct forecasts of non-events
  - More informative for rare events (Saito and Rehmsmeier 2015)
- Quality-decision threshold (new?)
  - α of user for whom forecast is “preferred” vs. d*

Datasets

- Theoretical Gaussian distributions
  - Focus on d'=1 with R=0.5, 1.0, 2.0
- US tornado warnings (Brooks and Correia 2018)
- Hidden slides
  - Storm Prediction Center forecasts (Hitchens and Brooks 2012)
  - Convection-allowing models updraft-helicity as forecast for severe (courtesy Burkely Gallo and Patrick Skinner)
    - Different thresholds at one time
    - Same threshold at different lead times

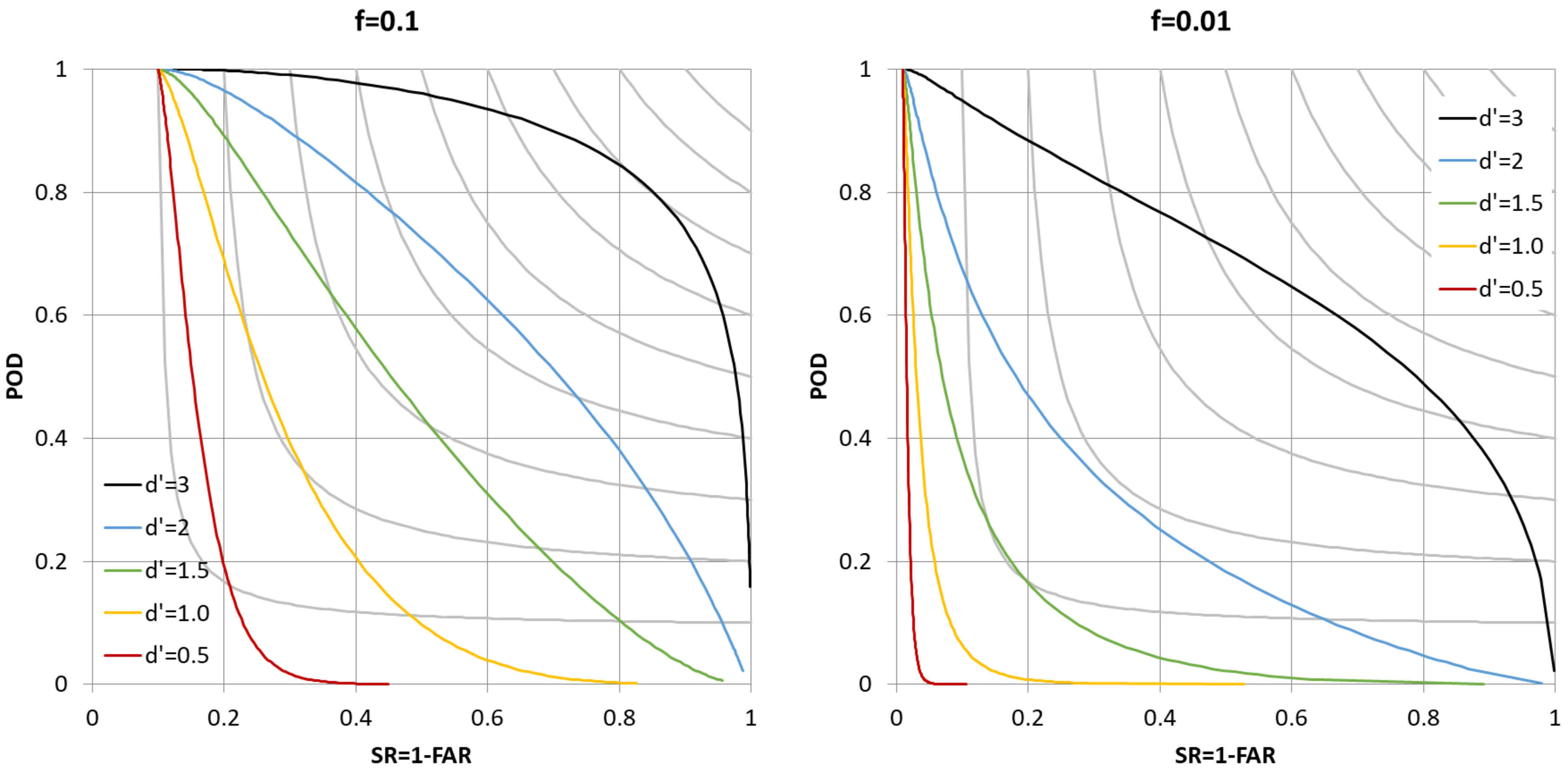
ROC diagram (R=1)

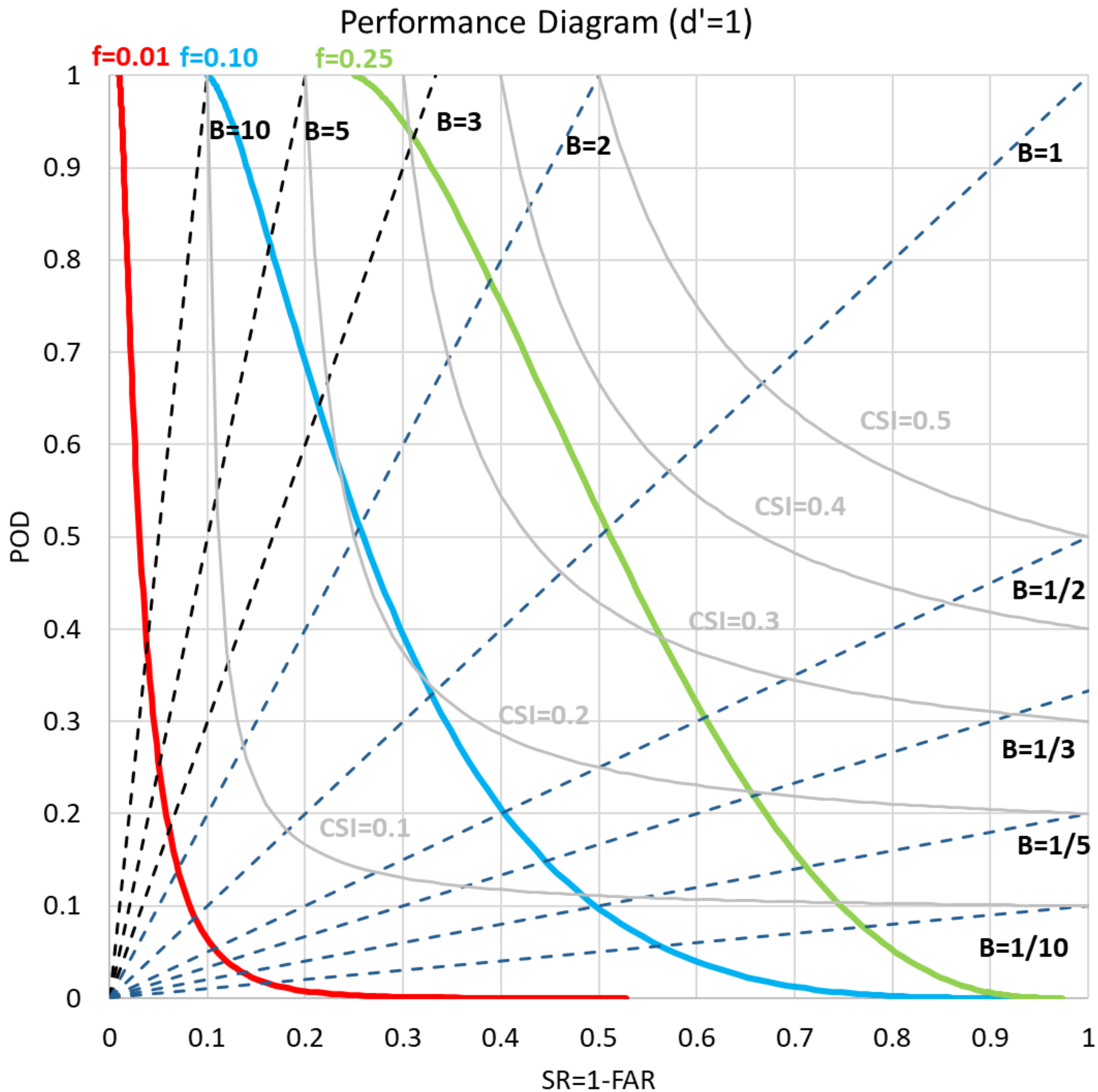


Impact of changing base rate

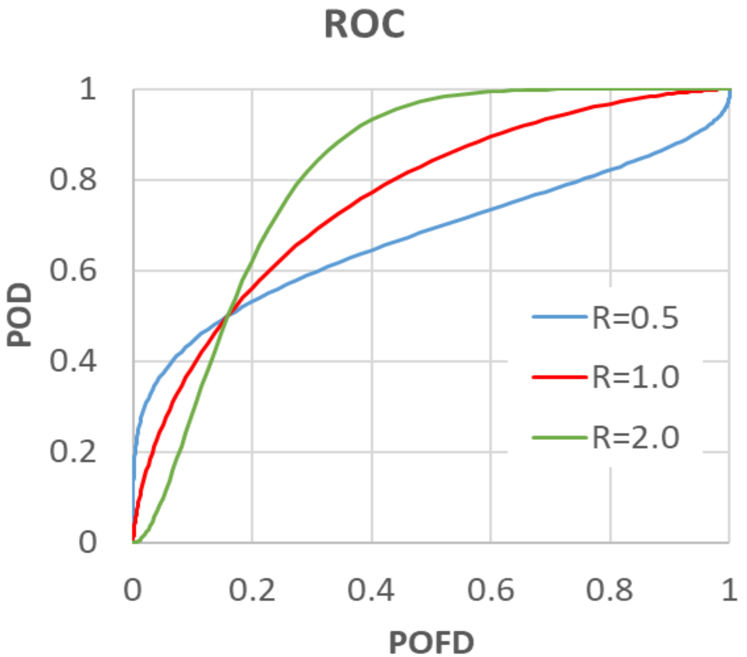
- Performance diagram
  - POD vs Success Ratio (1-FAR)
  - Has Bias, Critical Success Index information
  - Success Ratio is probability that event is “yes” if forecast is “yes”

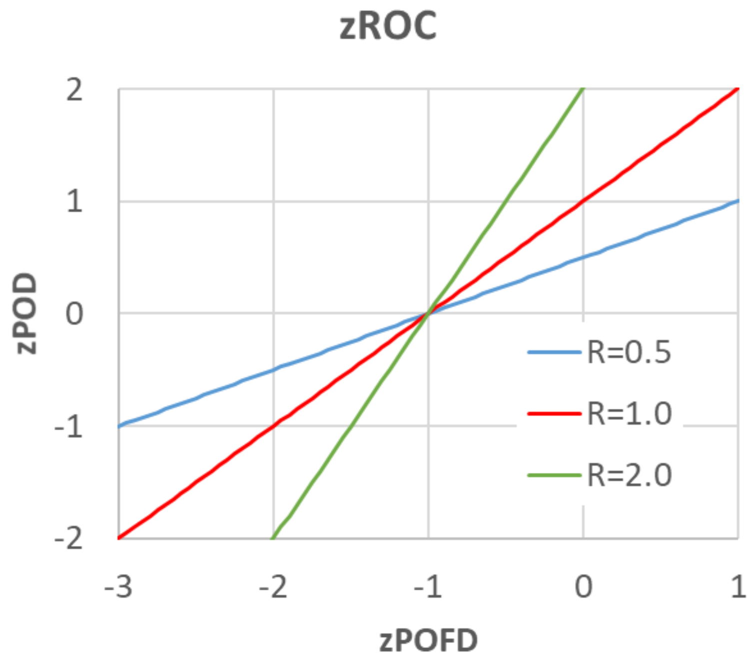
Performance diagrams




Performance Diagram (d'=1)
f=0.01
f=0.10
f=0.25
B=10
B=5
B=3
B=2
B=1
B=1/2
B=1/3
B=1/5
B=1/10
CSI=0.5
CSI=0.4
CSI=0.3
CSI=0.2
CSI=0.1
POD
SR=1-FAR
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

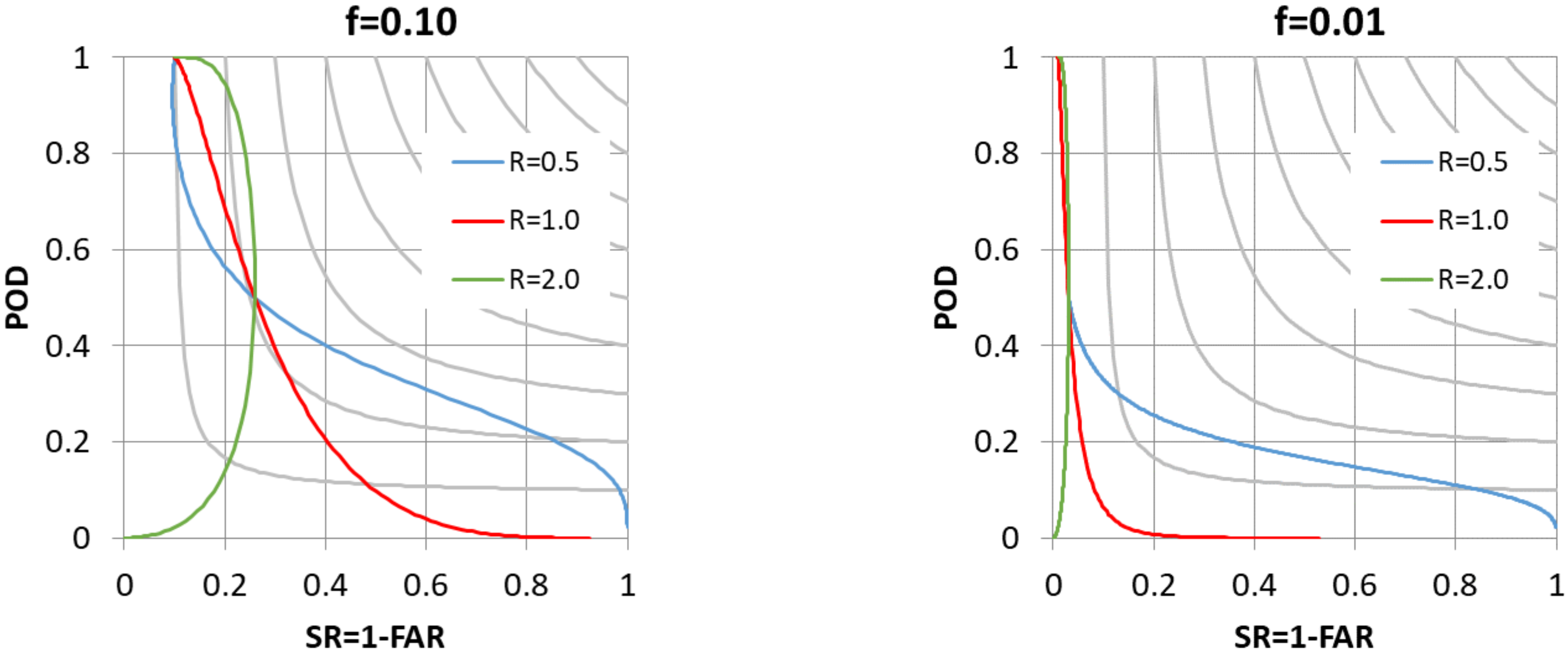
What if R≠1? (d’=1)





- Independent of base rate

Performance diagrams (d'=1)

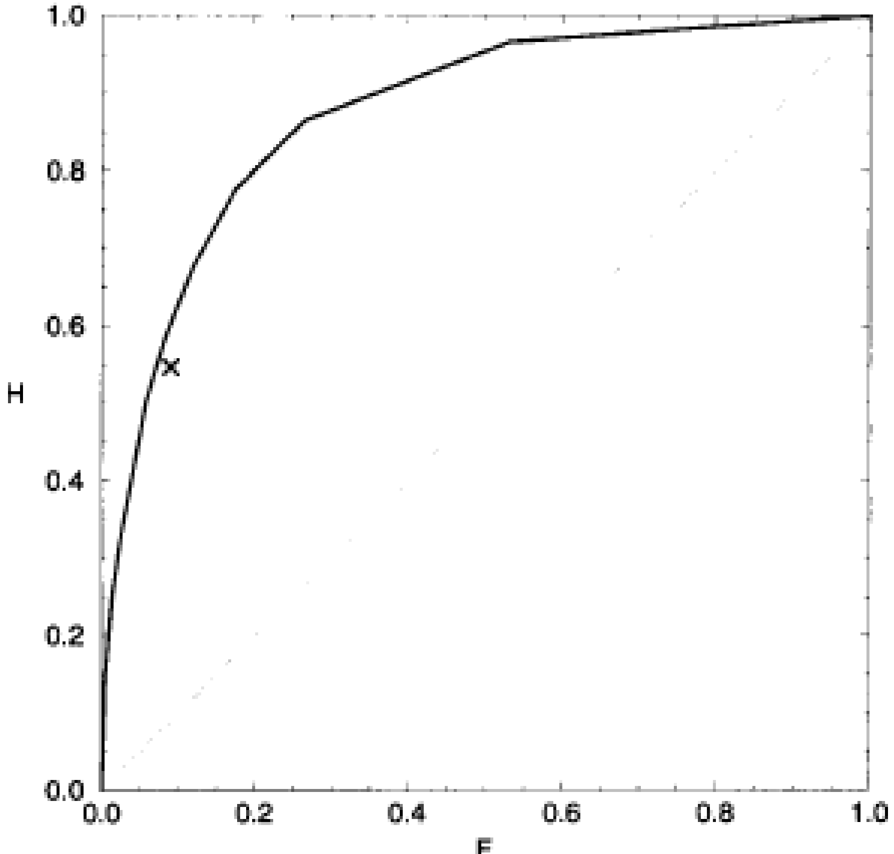


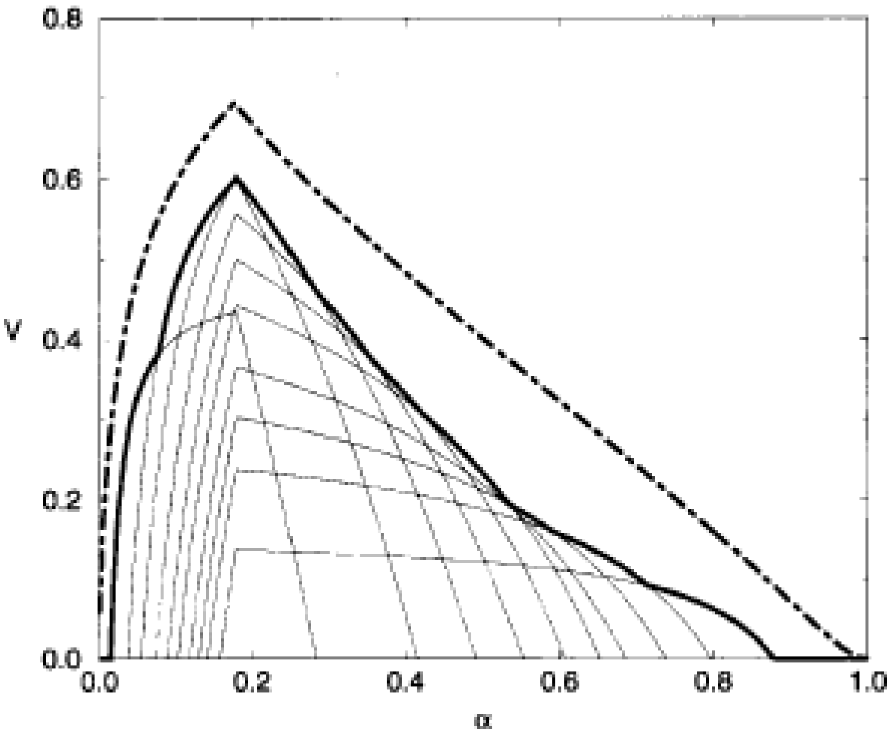
- Depends on base rate

Relating quality and relative value

- Richardson (2000)-cost-loss problem and relative value
  - Focused on probabilistic vs deterministic forecasts and impact of ensemble size

Richardson (2000)





- Relative value between base rate/perfect (α=cost loss)

Relating quality and relative value

- Richardson (2000)-cost-loss problem and relative value
  - Focused on probabilistic vs deterministic forecasts and impact of ensemble size
- Drummond and Holte (2006)
  - Combined base rate and costs of errors
  - Comparing different systems

Cost curves (Drummond and Holte 2006)

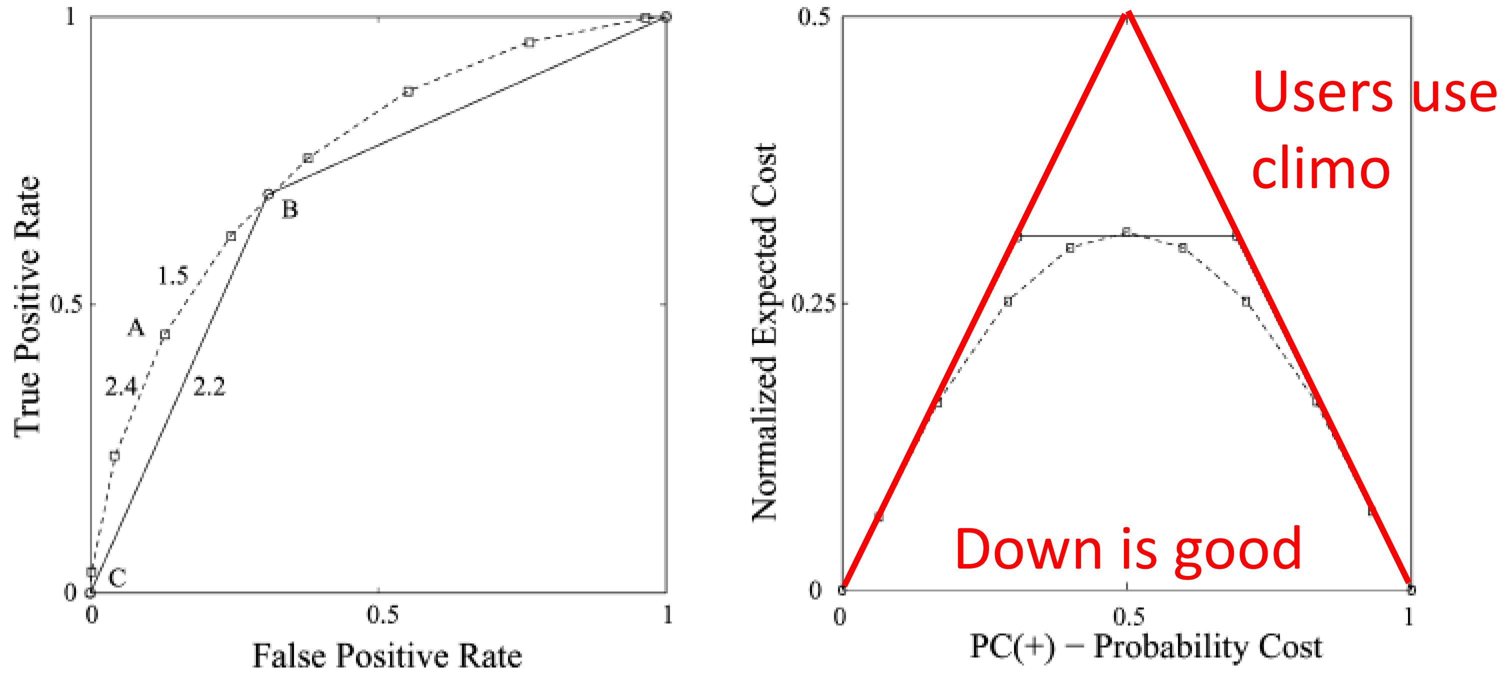


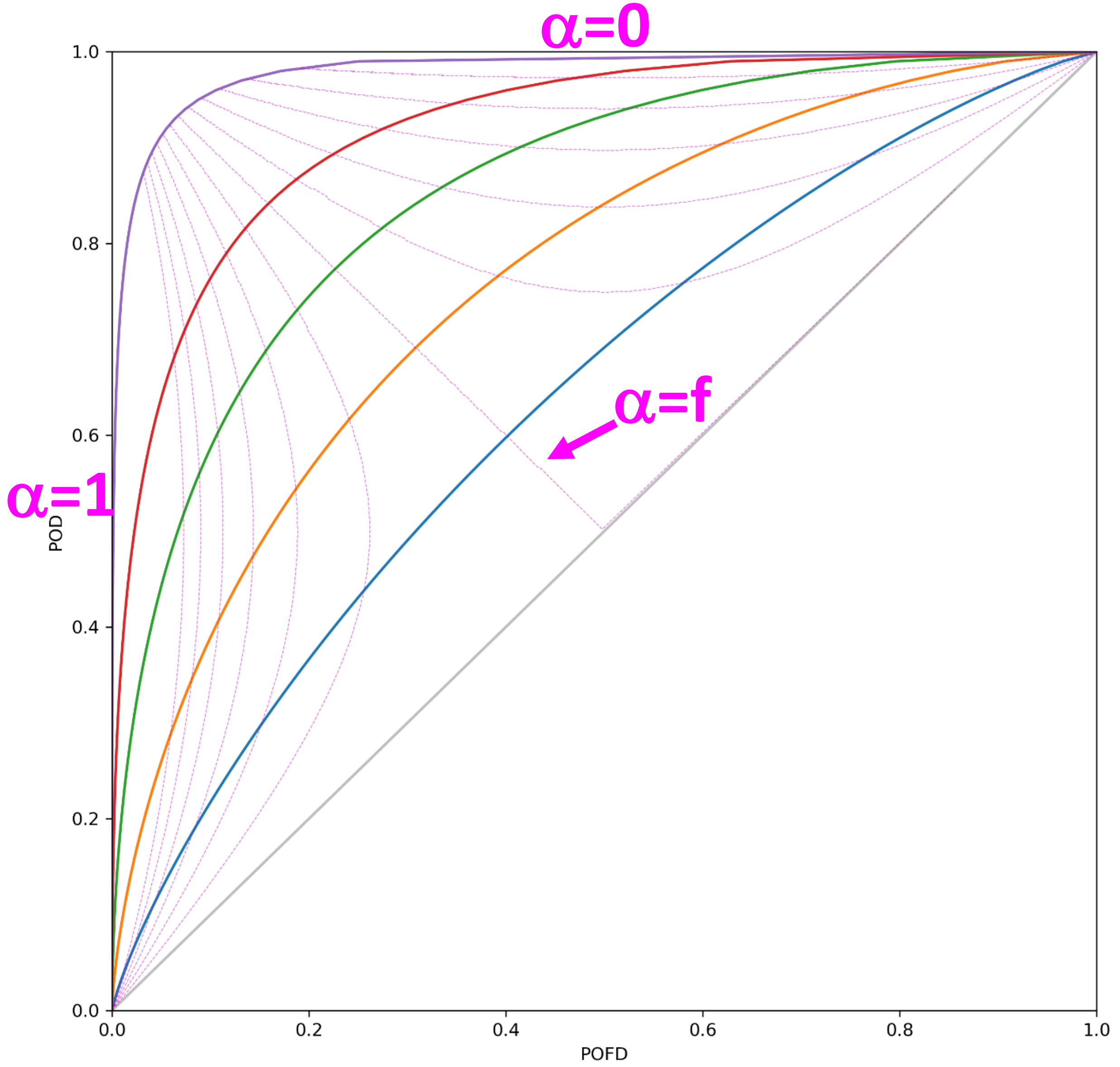
Fig. 12 (a) Two ROC curves whose performance is to be compared — (b) Corresponding cost curves

- PC(+)=p(y)*Cost(miss)/[p(y)*Cost(miss)+p(n)*cost(FA)]
- "Bidirectional point-line duality"!

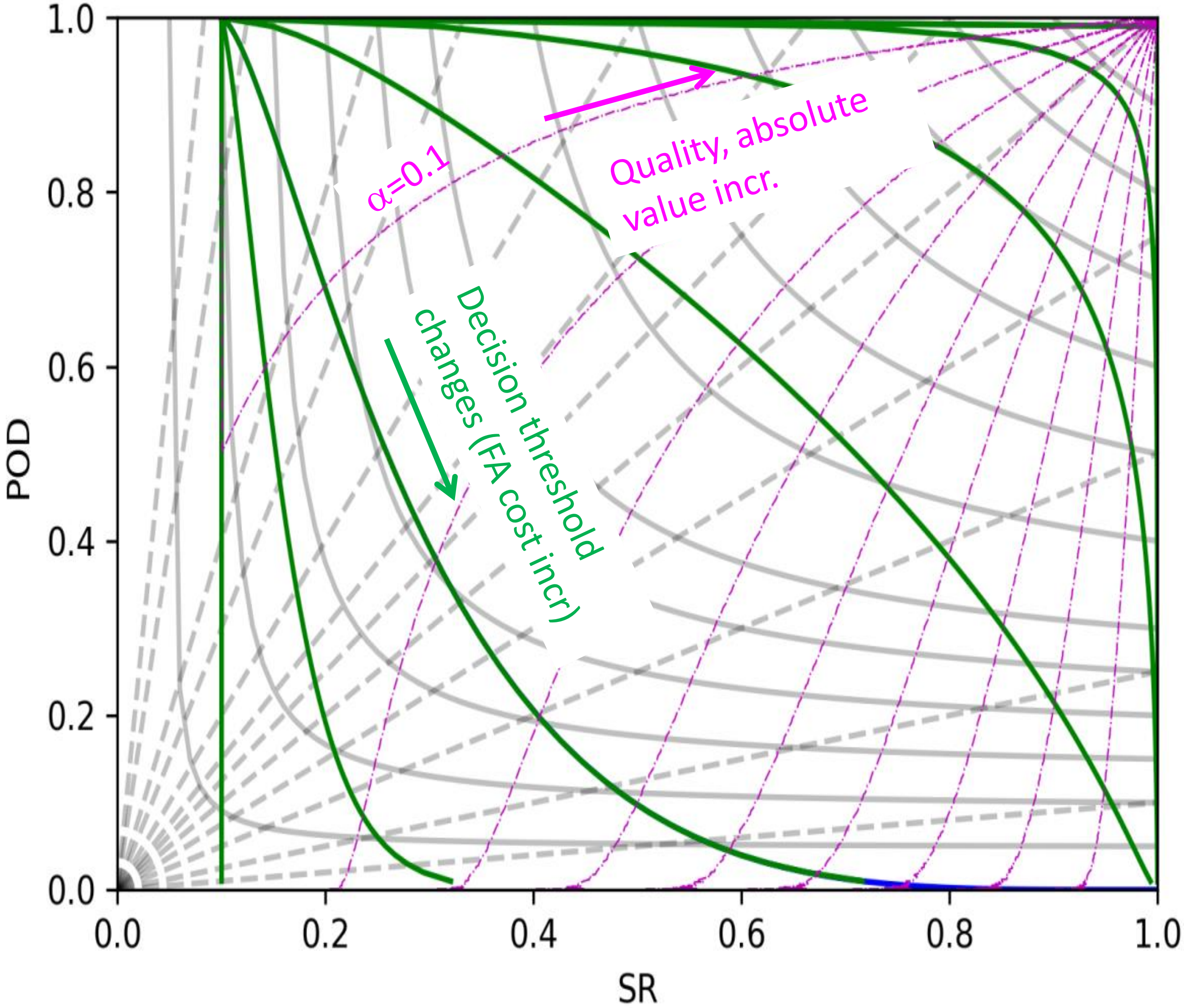
Finding Value

- Wandishin and Brooks (2002) show how to find relative value of forecasts in terms of POD, POFD, f, and α
- Implied α of system: Move along d’ curve and finding combo of POD and POFD associated with it
  - Cost associated with false alarm increases with α
- α between DFR and SR find value (operating range)
  - Low d* cut-off if $R \neq 1$ when users prefer “climo” forecast

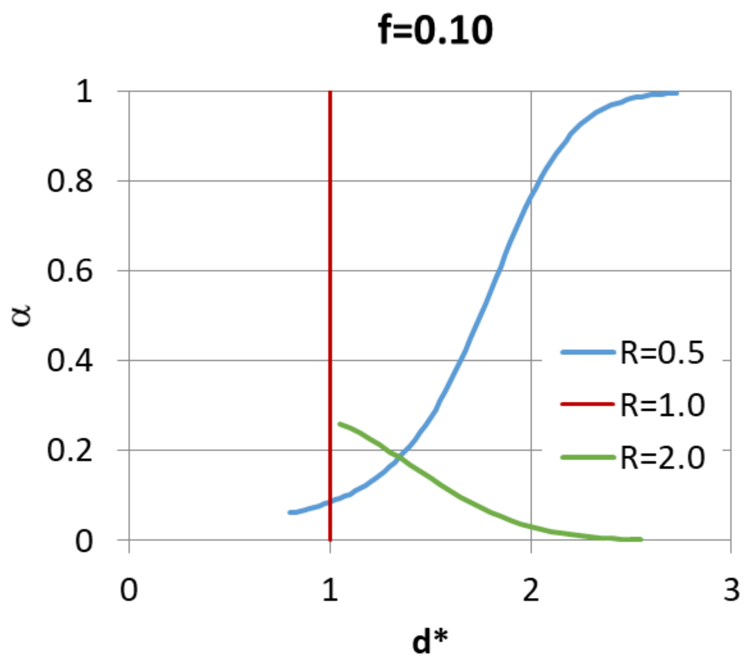
What α looks like on a ROC diagram

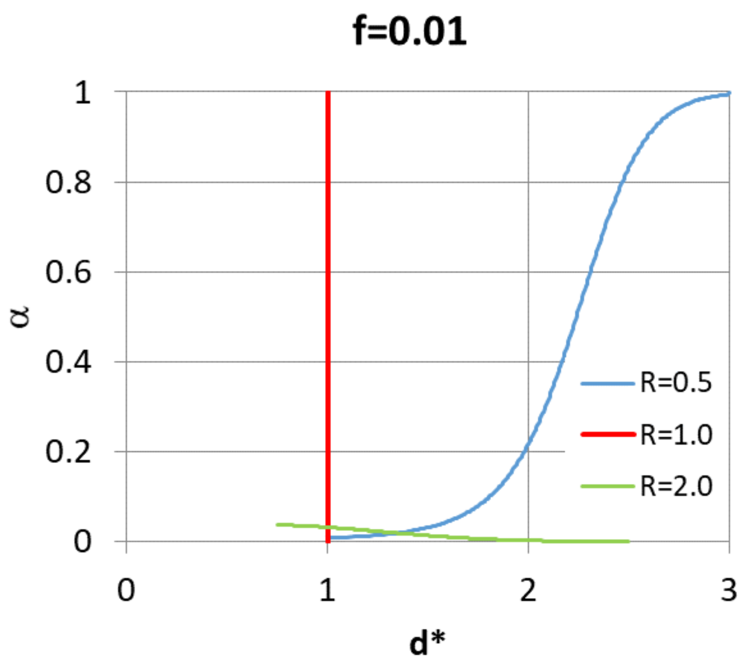


What α looks like on performance (R=1)



Quality-Decision Threshold



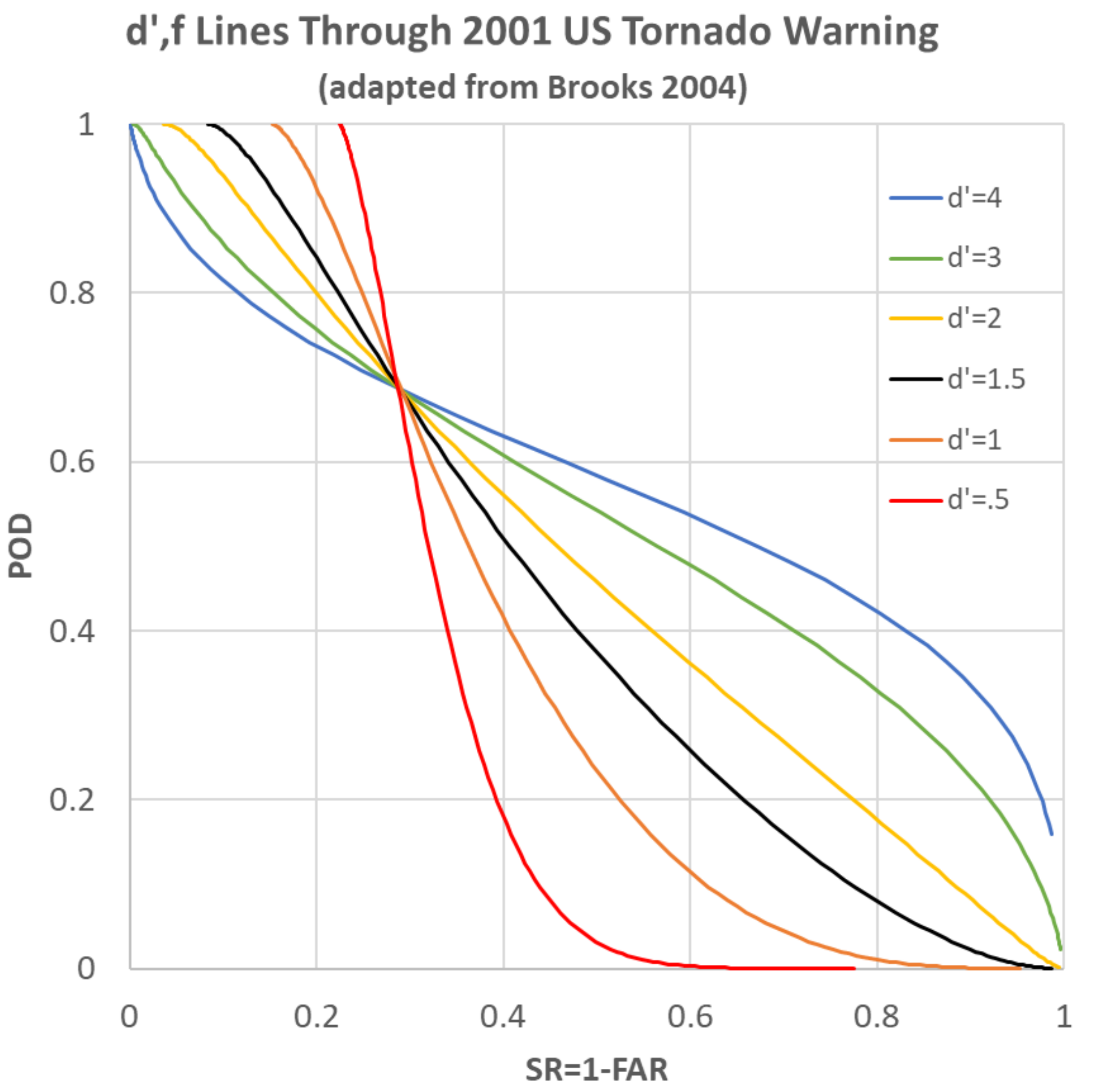


- Low-d* cut-off: all users prefer base rate forecasts
- “Non-vertical” QDT seen in CAM forecasts (hidden slide)

Looking at real forecasts

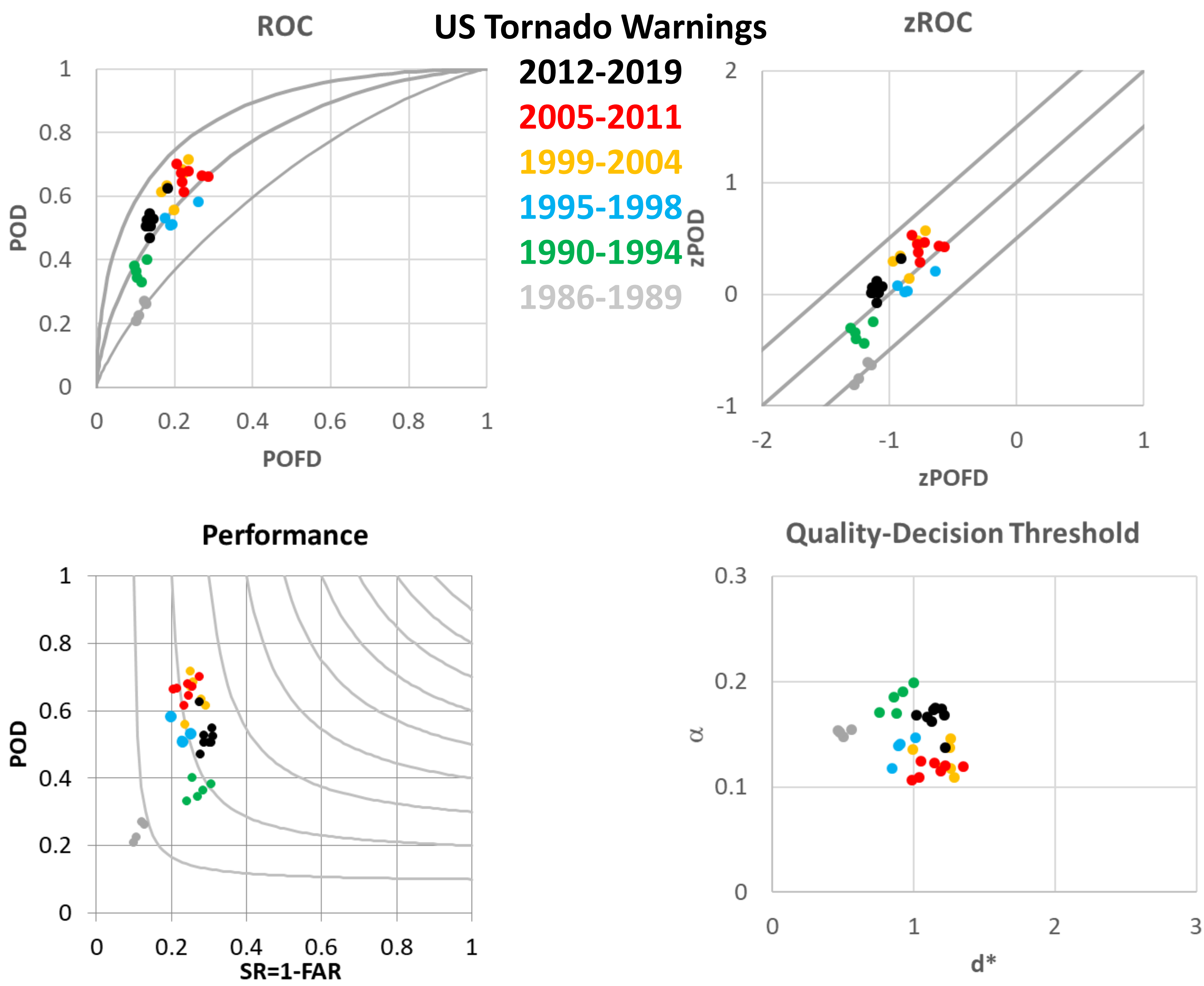
- Problem with correct forecasts of non-events
- Estimate either f (base rate of problem) or d in 2x2 table
  - Eliminate "easy" correct nulls-increases apparent f
  - Forecasting tornado vs. no storm or vs. severe non-tornadic?
  - High-res model-regions with clearly no threat?
- Ambiguity between f and d' has quantitative issues, but not qualitative
  - As f gets larger, d' gets smaller, QDT curves move up and to the left
-

Ambiguity of d' and f for real systems

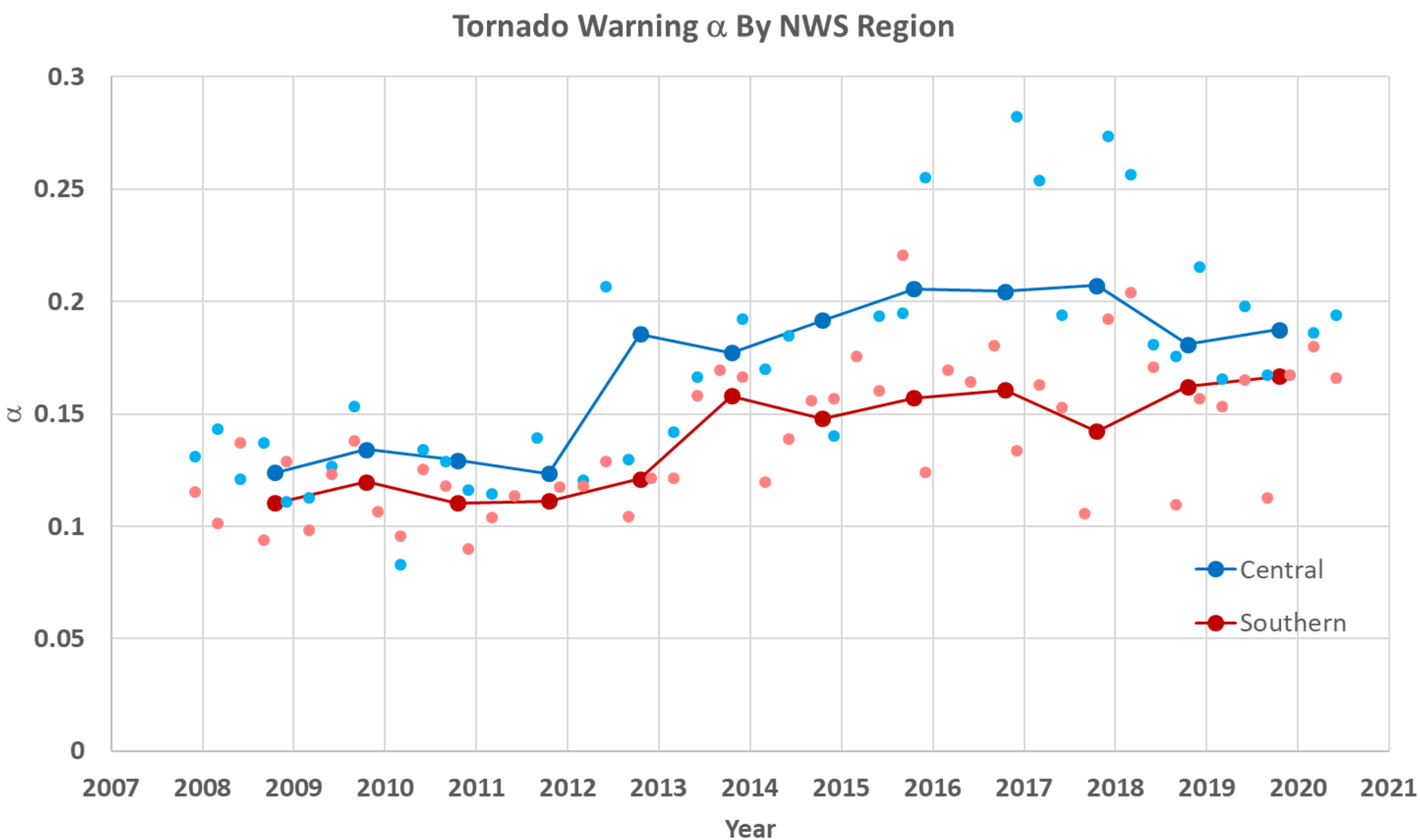


Looking at real forecasts

- Problem with correct forecasts of non-events
- Estimate either f (base rate of problem) or d in 2x2 table
  - Eliminate "easy" correct nulls-increases apparent f
  - Forecasting tornado vs. no storm or vs. severe non-tornadic?
  - High-res model-regions with clearly no threat?
- Ambiguity between f and d' has quantitative issues, but not qualitative
  - As f gets larger, d' gets smaller, QDT curves move up and to the left
- 4-panel figure for US annual tornado warning performance


US Tornado Warnings
2012-2019
2005-2011
1999-2004
1995-1998
1990-1994
1986-1989
ROC
POD
POFD
zROC
zPOD
zPOFD
Performance
POD
SR=1-FAR
Quality-Decision Threshold
α
d*

What happened in 2012/3?



Final thoughts

- Relationships between different metrics can be seen
  - Rare events: POD sensitive, FAR insensitive unless never forecast
  - For R=1, d' curves have max near bias=1 on performance
- Value curves can be drawn on ROC, performance
- Quality-decision threshold show changes in quality (d*) and the implied decision threshold (α)
- Monitoring can help identify changes in forecast system

References

- Brooks, H. E., 2004: Tornado warning performance in the past and future: A perspective from signal detection theory. *Bull. Amer. Meteor. Soc.*, **85**, 837–843.
- Brooks, H. E., and J. Correia Jr., 2018: Long-term performance metrics for National Weather Service tornado warnings. *Wea. Forecasting*, **33**, 1501-1511.
- Drummond, C., and R. C. Holte, 2006: Cost curves: An improved method for visualizing classifier performance. *Machine Learning,* **65**, 95-130
- Hitchens, N. M., and H. E. Brooks, 2012: Evaluation of the Storm Prediction Center's day 1 convective outlooks. *Wea. Forecasting*, **27**, 1580–1585.
- Mason, I. B., 1982: A model for assessment of weather forecasts. *Aust. Meteor. Mag*., **30**, 291–303.

References (cont.)

- Murphy, A. H., 1993: What is a good forecast? An essay on the nature of goodness in weather forecasting. *Wea. Forecasting*, **8**, 281–293.
- Richardson, D. S., 2000: Skill and relative economic value of the ECMWF ensemble pediction system. *Q. J. R. Meteorol. Soc*., **126**, 649–667.
- Roebber, P. J., and Bosart, L. E., 1996: The complex relationship between forecast skill and forecast value: a real-world analysis. *Wea. Forecasting*, **11**, 544–559.
- Roebber, P. J., 2009: Visualizing multiple measures of forecast quality. *Wea. Forecasting*, **24**, 601–608.
- Saito, T., and M. Rehmsmeier, 2015: The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLOS ONE*, DOI:10.1371/journal.pone.0118432
- Wandishin, M. S., and H. E. Brooks, 2002: On the relationship between Clayton's skill score and expected value for forecasts of binary events. *Meteor. Appl*., **9**, 455–459.