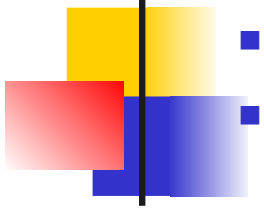
Verification of Categorical Forecasts – The Contingency Table

Laurence Wilson

laurence.Wilson@sympatico.ca

Co-chair, WMO Joint Working Group on Forecast Verification Research (JWGFVR)

Outline



- What defines an “event”
- Hits, misses, false alarms and correct negatives – the Contingency table
- Building the table
- Some relevant verification measures: Scores from the table and what they mean
- EXERCISE – Interpreting the table and scores

Resources

- Resources:
  - The EUMETCAL training site on verification – computer aided learning:
    - https://eumetcal.eu/links/
    - The website of the Joint Working Group on Forecast Verification Research:
    - *http://www.cawcr.gov.au/projects/**verification**/*
    - This contains definitions of all the basic scores and links to other sites for further information
  - Document “Verification of forecasts from the African SWFDPs” on the WMO website.

Why categorical?

- Inherently categorical
  - Precipitation yes or no
  - Precipitation type
  - Threshold accumulation
    - 0.5 mm? 0.2 mm?....
- User importance
  - Does the wind matter if it is less than 5 m/s?
  - Does it matter if 32 or 34 mm of precipitation fell?
  - Extremes…>50 mm rain in 24h….
  - High impact weather

What is truth? Some comments on observations

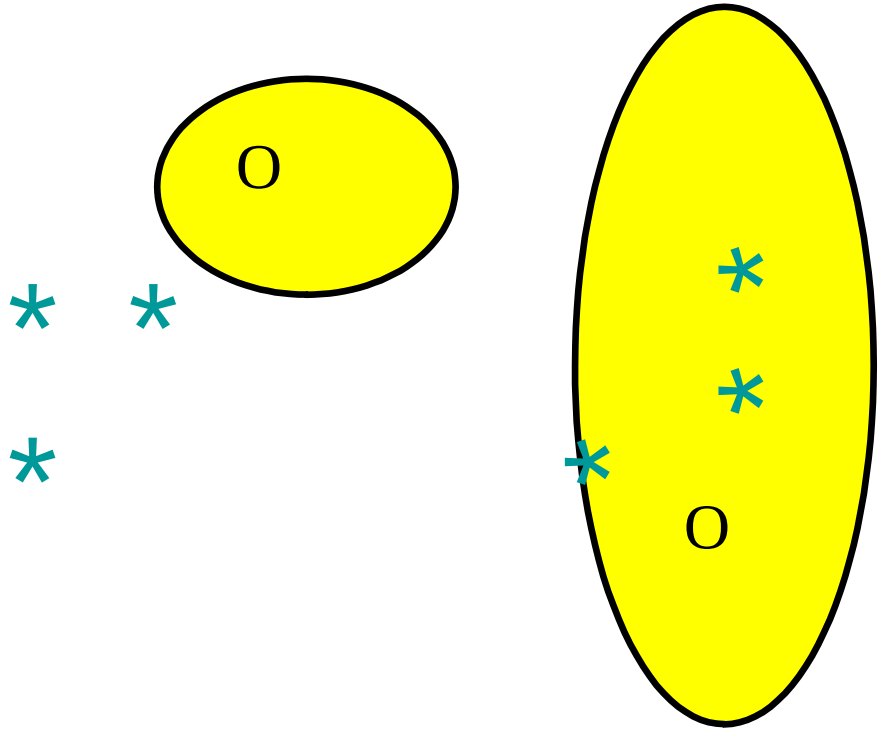
- Station observations
  - Valid at points – a sample of local weather
  - Generally accurate for the points they represent
  - BUT must be quality controlled
  - For verification, QC should be independent of models
- Satellite-derived precipitation estimates such as HE
  - Space and time coverage good if from geostationary
  - NOT representative of points – some averaging e.g. HE is about 12km. Limited by satellite footprint

What is the Event?

- For categorical and probabilistic forecasts, one must be clear about the “event” being forecast
  - Location or area for which forecast is valid
  - Time range over which it is valid
  - Definition of category
- And now, what is defined as a correct forecast?
  - The event is forecast, and is observed – anywhere in the area? Over some percentage of the area?
  - Scaling considerations

Verification of NMS warnings: What is the Event?

- Then, how to match observed “events” to forecast:
  - Location or area for which forecast is valid
  - Time range over which it is valid
  - Definition of category
- And now, what is defined as a correct forecast?
  - The event is forecast, and is observed – anywhere in the area? Over some percentage of the area?



Summary - Events

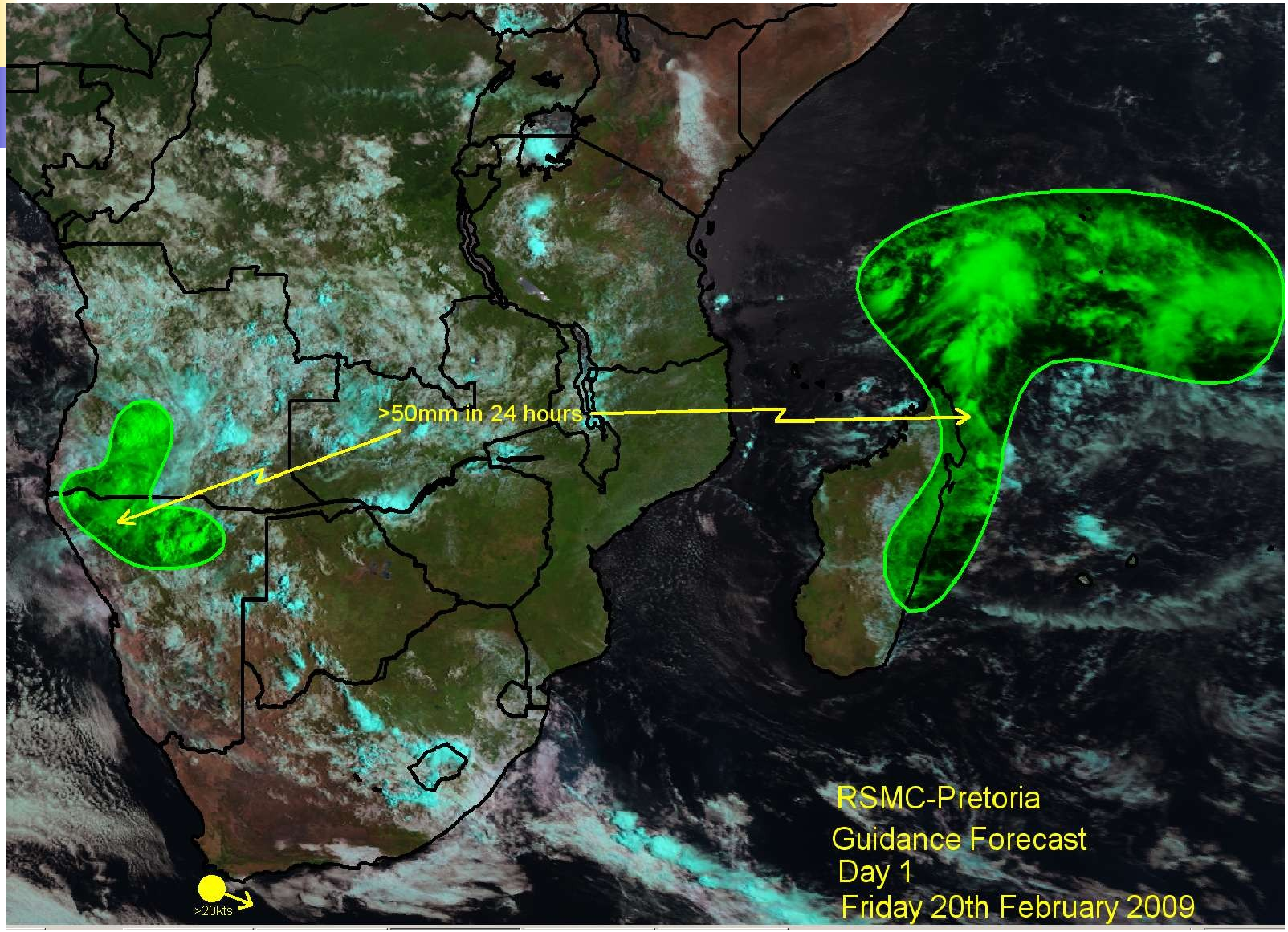
- Best if “events” are defined for similar time period and similar-sized areas
  - One day 24h
  - Fixed areas; should correspond to forecast areas and have at least one reporting stn.
  - Data density a problem
    - Best to avoid verification where there is no data.
  - Non-occurrence – no observation problem
- Observation – based reporting
  - The event is defined by the observation
  - Can therefore have both hits and false alarms inside a forecast severe weather area.
  - Observations outside a severe weather forecast area are misses
  - All observations lower than threshold value outside forecast threat areas are correct negatives

Preparation of the contingency table

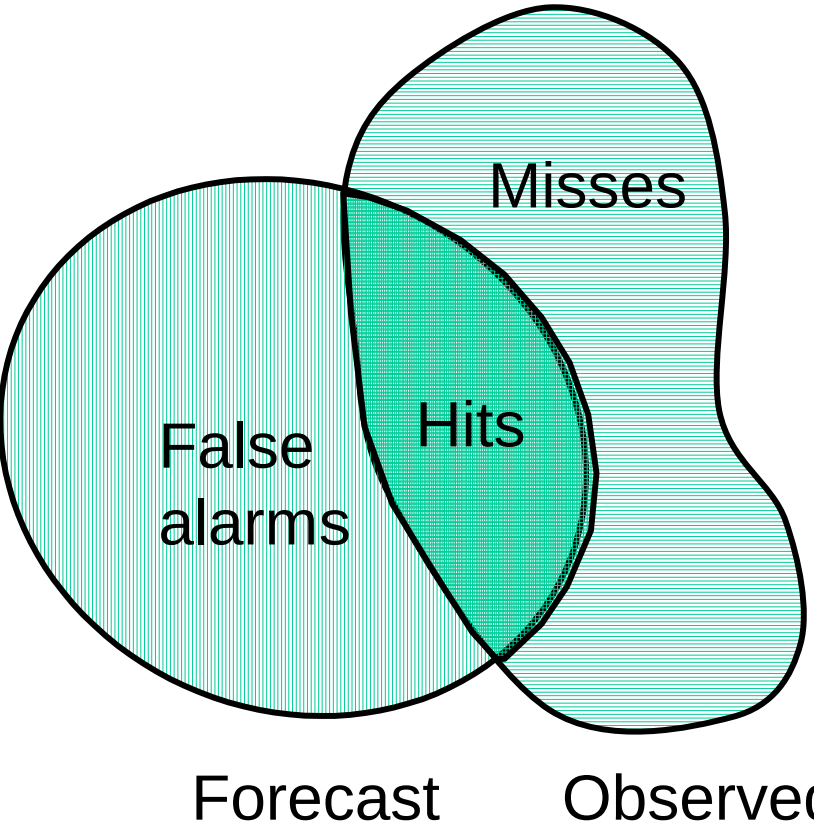
- Start with matched forecasts and observations
- Forecast event is precipitation >50 mm / 24 h Next day
- Count up the number of each of hits, false alarms, misses and correct negatives over the whole sample
- Enter them into the corresponding 4 boxes of the table.

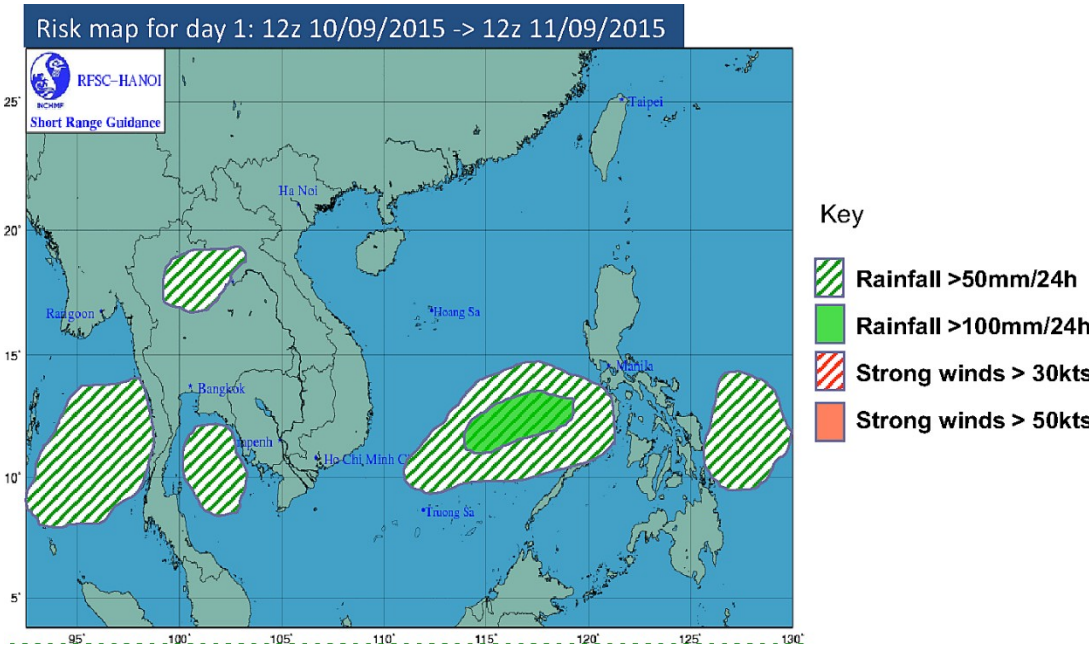
Day	Fcst to occur?	Observed ?
1	Yes	Yes
2	No	Yes
3	No	No
4	Yes	No
5	No	No
6	Yes	Yes
7	No	No
8	No	Yes
9	No	No

How do we verify this?



Spatial verification of RMSC products



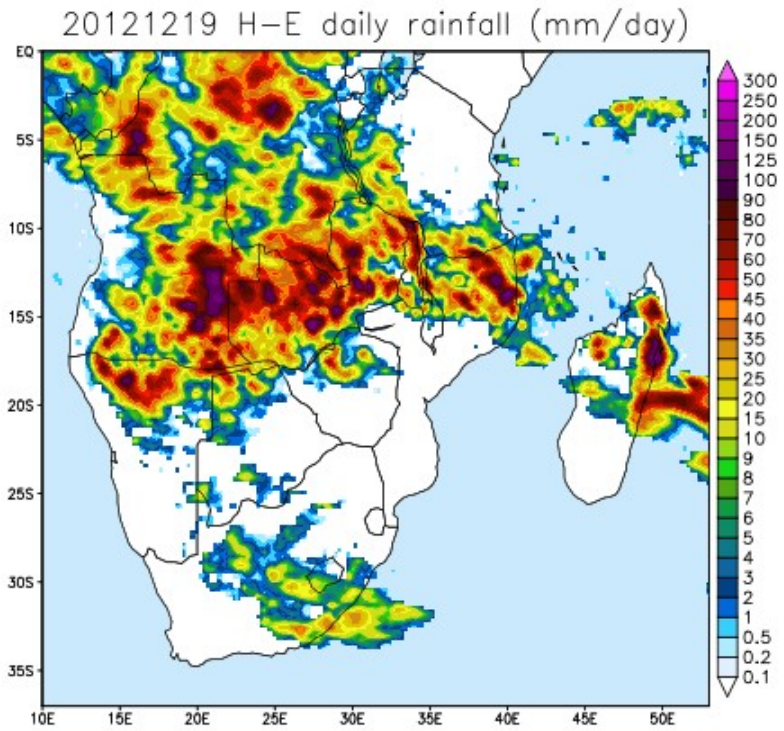


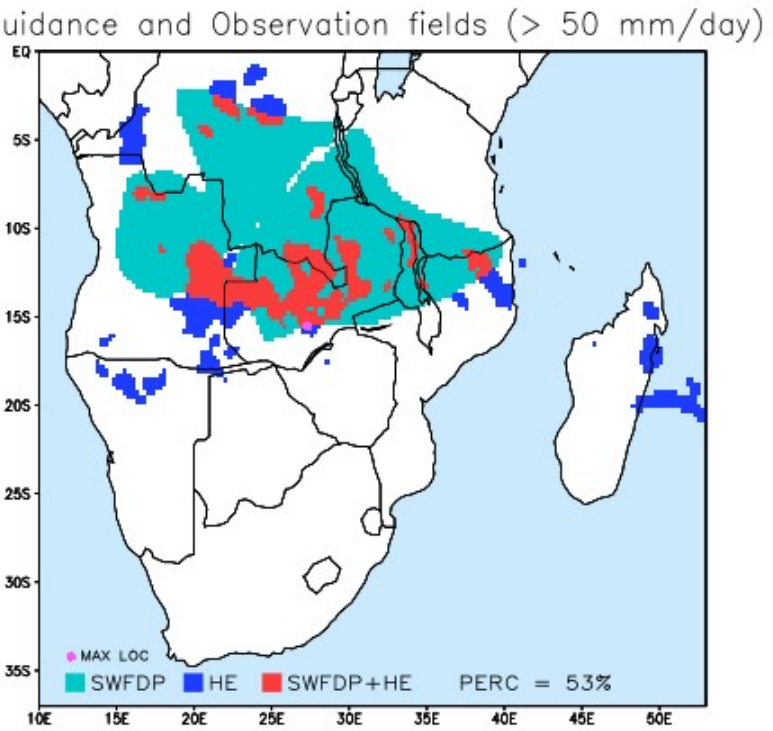
Spatial contingency table:

-Can accomplish IF one has quasi-continuous spatial observation data

-Stephanie's method

Verification of regional forecast map using HE





Verification statistics for 20121219 : Grid Size = 0.25° : Units = mm/day : n = 25777

	Guidance	H-E
Number of gridpoints >= 50 mm	3294	1243
Average Rain over domain	~	19.7012
>= 50 mm Rain Area (km²*10⁶)	2.05875	0.776875
Maximum Rainfall Observed (mm)	~	151.124

Categorical Forecasts	
Frequency Bias	2.65004
Probability of Detection	0.526146
False Alarm Ratio	0.801457
Hansen & Kuipers Score	0.418541
Equitable threat score	0.132959
Spatial Correlation	0.264835

	OBSERVATION >=50	OBSERVATION <50
GUIDANCE >=50	654	2640
GUIDANCE <50	589	21894

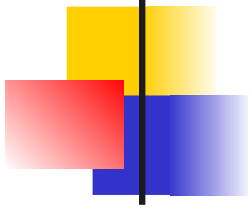
Extreme Events Verification	
Extreme Dependency Score	0.650434
Symmetric Extreme Dependency Score	0.385181
Extremal Dependency Index	0.552717
Symmetric Extremal Dependency Index	0.59486
(**Ferro and Stephenson, 2011***)	

http://rsmc.weathersa.co.za/RSMC/index.php
Format based on IPWG verification output

WMO

South African Weather Service

The contingency Table



Observations

Forecasts	Yes	No	
Yes	HITS	FALSE ALARMS	Total Events Forecast
No	MISSED EVENTS	CORRECT NEGATIVES	Total non-events Forecast
	Total Events Observed	Total Non-Events Observed	Sample size

Contingency tables

Forecasts \ Observations			
	HITS a	FALSE ALARMS b	Total Events Forecast a+b
	MISSED EVENTS c	CORRECT NEGATIVES d	Total non-events Forecast c+d
	Total Events Observed a+c	Total Non-Events Observed b+d	Sample size T=a+b+c+d

$$PoD = \frac{a}{a+c}$$

range: 0 to 1
best score = 1

$$FAR = \frac{b}{(a+b)}$$

range: 0 to 1
best score = 0

Characteristics:

- PoD= “Prefigurance” or “probability of detection”, “hit rate”
  - Sensitive only to missed events, not false alarms
  - Can always be increased by overforecasting rare events
- FAR= “False alarm ratio”
  - Sensitive only to false alarms, not missed events
  - Can always be improved by underforecasting rare events

Contingency tables

Forecasts \ Observations			
	HITS a	FALSE ALARMS b	Total Events Forecast a+b
	MISSED EVENTS c	CORRECT NEGATIVES d	Total non-events Forecast c+d
	Total Events Observed a+c	Total Non-Events Observed b+d	Sample size T=a+b+c+d

$$PAG = \frac{a}{a+b}$$

range: 0 to 1
best score = 1

$$Bias_{frequency} = \frac{a+b}{a+c}$$

best score = 1

Characteristics:

- PAG= “Post agreement”
  - PAG= (1-FAR), and has the same characteristics
- Bias: This is frequency bias, indicates whether the forecast distribution is similar to the observed distribution of the categories (Reliability)

What's wrong with PC - % correct?
The Finley Affair (1884)

		Observed		
		tornado	no tornado	Total
Forecast	tornado	**28**	**72**	**100**
	no tornado	**23**	**2680**	**2703**
Total		**51**	**2752**	**2803**

% correct = (28+2680)/2803 =96.6%; No tornado forecast: (2752)/2803 =98.2%!

Contingency tables

Forecasts \ Observations			
	HITS a	FALSE ALARMS b	Total Events Forecast a+b
	MISSED EVENTS c	CORRECT NEGATIVES d	Total non-events Forecast c+d
	Total Events Observed a+c	Total Non-Events Observed b+d	Sample size T=a+b+c+d

$$CSI = \frac{a}{a+b+c}; \frac{d}{b+c+d}$$

range: 0 to 1
best score = 1

Characteristics:

- Better known as the Threat Score
- Sensitive to both false alarms and missed events; a more balanced measure than either PoD or FAR
- ETS = Equitable threat score is the TS adjusted for number correct by chance

Contingency tables

Forecasts \ Observations			
	HITS a	FALSE ALARMS b	Total Events Forecast a+b
	MISSED EVENTS c	CORRECT NEGATIVES d	Total non-events Forecast c+d
	Total Events Observed a+c	Total Non-Events Observed b+d	Sample size T=a+b+c+d

$$HSS = \frac{(a+d) - \frac{(a+b)(a+c)+(c+d)(b+d)}{T}}{T - \frac{(a+b)(a+c)+(c+d)(b+d)}{T}}$$

range: negative value to 1
best score = 1

Characteristics:

- A skill score against chance (as shown)
- Easy to show positive values
- Better to use climatology or persistence
  - needs another table

$$ETS = \frac{a - \frac{(a+b)(a+c)}{T}}{a+b+c - \frac{(a+b)(a+c)}{T}}$$

Contingency tables

Observations

Forecasts

HITS a	FALSE ALARMS b	Total Events Forecast a+b
MISSED EVENTS c	CORRECT NEGATIVES d	Total non-events Forecast c+d
Total Events Observed a+c	Total Non-Events Observed b+d	Sample size T=a+b+c+d

$$HR = \frac{a}{a+c}$$

range: 0 to 1
best score = 1

$$FA = \frac{b}{(b+d)}$$

best score = 0

$$KSS = HR - FA$$

Characteristics:

- Hit Rate (HR) is the same as the PoD and has the same characteristics
- False alarm RATE. This is different from the false alarm ratio.
- These two are used together in the Hanssen-Kuipers (Pierce, True skill statistic) score, and in the ROC, and are best used in comparison.

Verification of extreme, high-impact weather

- **EDS – EDI – SEDS - SEDI Novelty categorical measures!**

Standard scores tend to zero for rare events

Event forecast	Event observed		
	Yes	No	Marginal total
Yes	a	b	a + b
No	c	d	c + d
Marginal total	a + c	b + d	a + b + c + d =n

H = a / (a+c), hit rate

F = b / (b+d), false alarm rate

p = (a+c) / n, base rate

q = (a+b) / n, relative frequency of forecasted events

$$\mathrm{EDS} = \frac{\log p - \log H}{\log p + \log H} \qquad \mathrm{SEDS} = \frac{\log q - \log H}{\log p + \log H}$$

Ferro & Stephenson, 2011: Improved verification measures for deterministic forecasts of rare, binary events. *Wea. and Forecasting*

Base rate independence Functions of H and F

$$\mathrm{EDI} = \frac{\log F - \log H}{\log F + \log H}$$

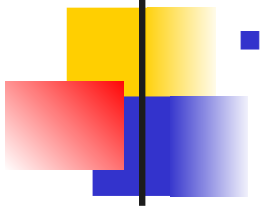
Extremal Dependency Index - EDI

Symmetric Extremal Dependency Index - SEDI

$$\mathrm{SEDI} = \frac{\log F - \log H - \log(1-F) + \log(1-H)}{\log F + \log H + \log(1-F) + \log(1-H)}$$

JWGFVR

Comments on the extreme dependency family



- EDS now discredited
  - Sensitive to base rate
  - NOT sensitive to false alarms
- SEDS
  - Weakly sensitive to base rate, but useful
  - Useful to forecasters because uses the forecast frequency
- EDI
  - User-oriented, function of HR and FA like HK and ROC
  - Absolutely independent of base rate
- SEDI
  - Like EDI, but has additional property of symmetry; not necessarily important for our purposes

Example - Madagascar

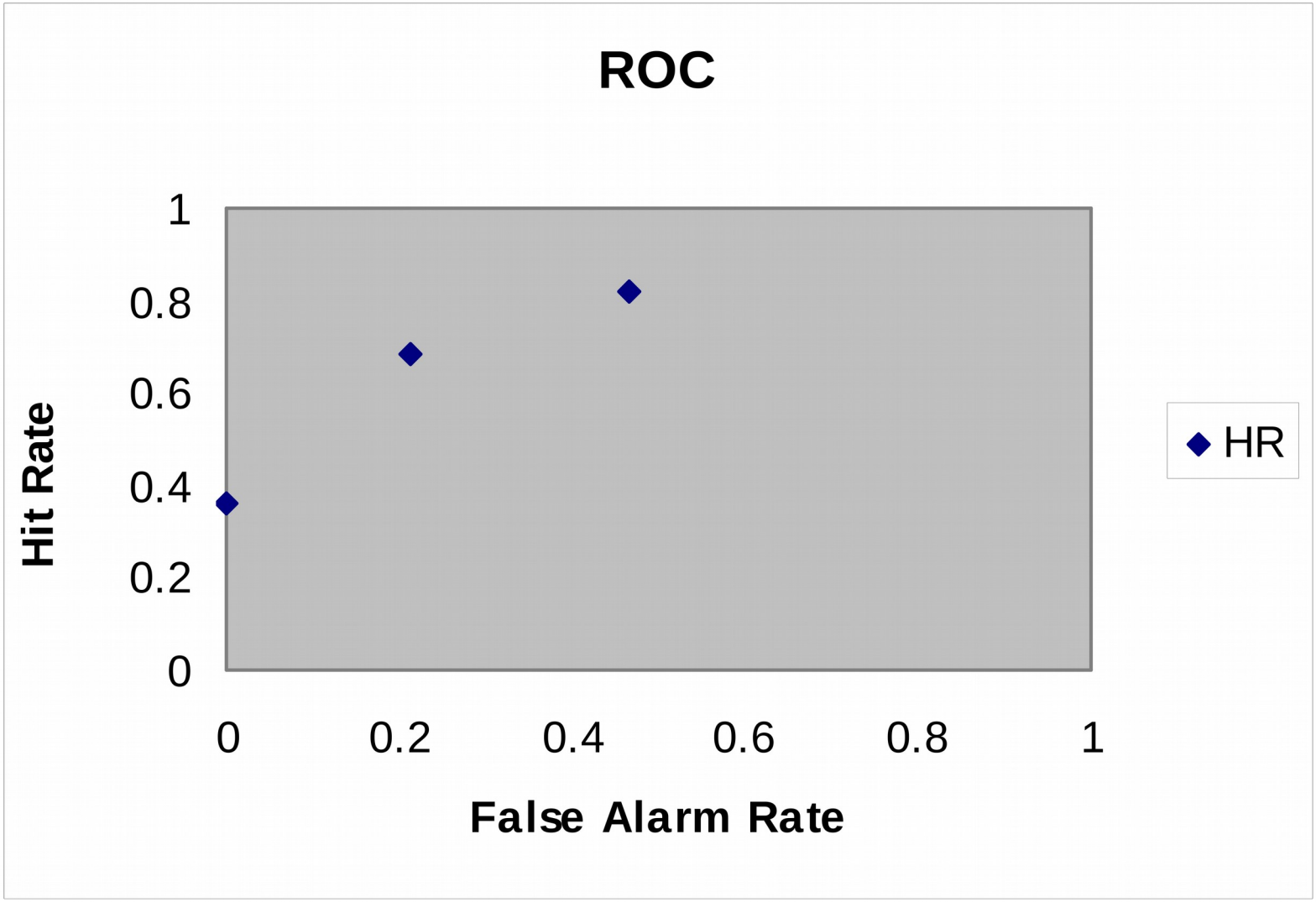
- 78 Cases
- Separate tables assuming low, medium, high risk as thresholds
- Can plot the hit rate vs the false alarm RATE = FA/total obs no

Low	Obs yes	Obs no	Totals
Fcst yes	18	26	44
Fcst no	4	30	34
Totals	22	56	78

Med	Obs yes	Obs no	Totals
Fcst yes	15	12	27
Fcst no	7	44	51
Totals	22	56	78

High	Obs yes	Obs no	Totals
Fcst yes	8	0	8
Fcst no	14	56	70
Totals	22	56	78

Example (contd)



Exercises

- 1. Three model comparison
  - 2014 data, ECMWF, GSM (Japan) and GFS (USA)
  - 6 SE Asia stations
  - Same observation dataset for all models
  - Contingency table for thresholds 0.5 mm to 50 mm / 24h
  - Using Excel
- 2. ECMWF 2016 dataset for 3 different stations