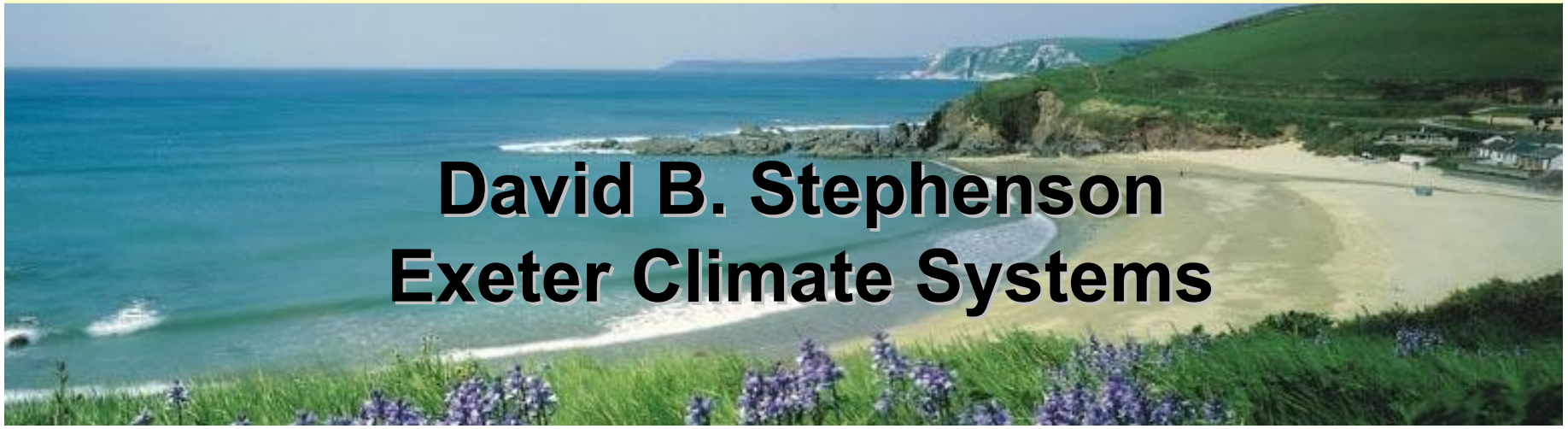
Verification of Weather Warnings

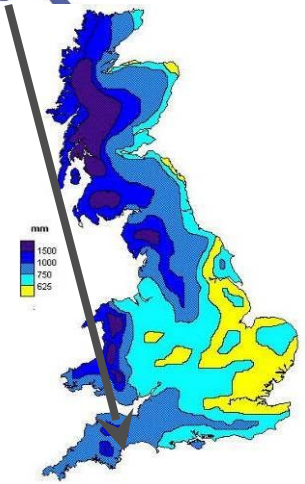
Did the boy cry wolf or was it just a sheep?



David B. Stephenson
Exeter Climate Systems

Jolliffe, Clive Wilson, Michael Sharpe,
Hewson, and Marion Mittermaier.

nks also to Harold Brooks, Chris Ferro,
Glahn, Martin Goeber and Dan Wilks.



Invited talk at 4th International Verification Methods workshop
Helsinki, Finland, 7-10 June 2009

Did the boy cry wolf or was it just a sheep?

Too many false alarms → loss of credibility → warning ignored

Hits, false alarms, and misses

A HIT
The boy cries wolf
and a wolf appears

A FALSE ALARM
The boy cries wolf
but a sheep appears

A MISS
The boy cries sheep
but a wolf appears!

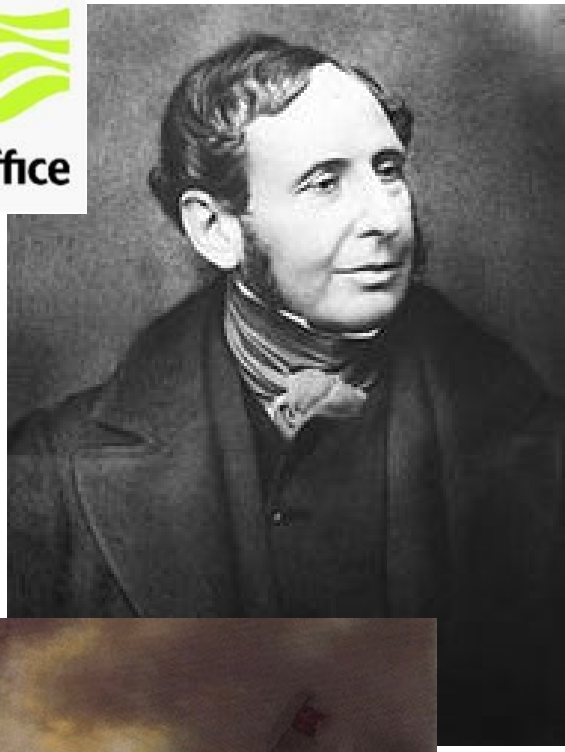
Storm warnings: birth of the UK Met Office

- **Admiral Robert Fitzroy CB FRS, 1805-1865 appointed in 1854 as chief to deal with the collection of weather data at sea, with the title of *Meteorological Statist to the Board of Trade* and a staff of three. This was the forerunner of the modern Meteorological Office;**

- **A terrible storm in 1859 that caused the loss of the *Royal Charter* inspired Fitzroy to develop charts to allow predictions to be made, which he called *"forecasting the weather"*, thus coining the term weather forecast.**

- **15 land stations were established to use the new telegraph to transmit to him daily reports of weather at set times. The first daily weather forecasts were published in *The Times* in 1860, and in the following year a system was introduced of hoisting storm warning cones at the principal ports when a gale was expected.**

- **Many fishing fleet owners objected to gale warnings, requiring that fleets not leave the ports and under this pressure, Fitzroy's system was abandoned for a short time after his death.**



Warnings issued by the Met Office

- National Severe Weather Warnings
- Extreme Rainfall Warnings **NEW**
- Marine warnings: coastal strong winds, gale warnings, storm warnings
- Heat-health warnings
- Open road warnings
- Defence warnings

Warnings: definition and features

Warning = deterministic forecast of severe weather

- Easy-to-use decision support tool;
- Simpler to communicate than probabilistic forecasts;
- Decision-thresholds chosen by the forecaster

See Murphy, A., 1991: "Probabilities, Odds, and Forecasts of Rare Events", Weather and Forecasting, Vol. 6, 302-307;

- Widely issued by weather services;
- Not widely discussed in the verification literature

Example: Inshore water forecast

Lyme Regis to Lands End including the Isles of Scilly

Coastal strong wind warning 03:26 Thu 08 Nov 18:00

Westerly winds, veering northwesterly later, will increase

this morning to reach Force 6 or 7 at times.

Inshore waters forecast 24 hour forecast:

0600 Thu 08 Nov 0600 Fri 09 Nov

Wind Northwest, backing west for a time, 4 or 5 inc. 5 to 7.

Sea state Slight increasing moderate, occasionally rough.

Warnings as a set of three times

Rattray Head to Berwick in the first three days of January 2007:

Y0	M0	D0 HHMM0	Y1	M1	D1 HHMM1	Y2	M2	D2 HHMM2
2007	01	01 0432	2007	01	01 0432	2007	01	01 1800
2007	01	01 1624	2007	01	01 1624	2007	01	02 0600
2007	01	02 0409	2007	01	02 0409	2007	01	02 1800
2007	01	02 1616	2007	01	02 1616	2007	01	03 0600
2007	01	03 0429	2007	01	03 0429	2007	01	03 1800
2007	01	03 1451	2007	01	03 1451	2007	01	03 1800
2007	01	03 1512	2007	01	03 1512	2007	01	04 0600

year: month: day: hours: minutes.

Warning is a set of three times (T0,T1,T2):

T0 = time the warning is issued

T1 = start time of warning period

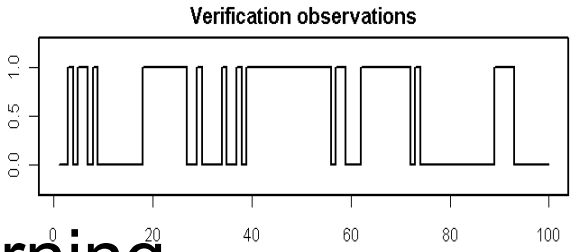
T2 = end time of warning period

T1-T0 is the lead time (can be 0 or even <0!)

Warnings as a binary function of time

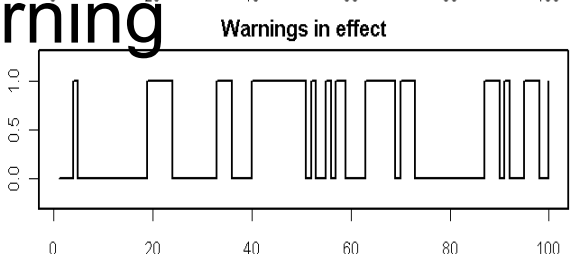
No observed event, obs event

Y(t) = 0 or 1

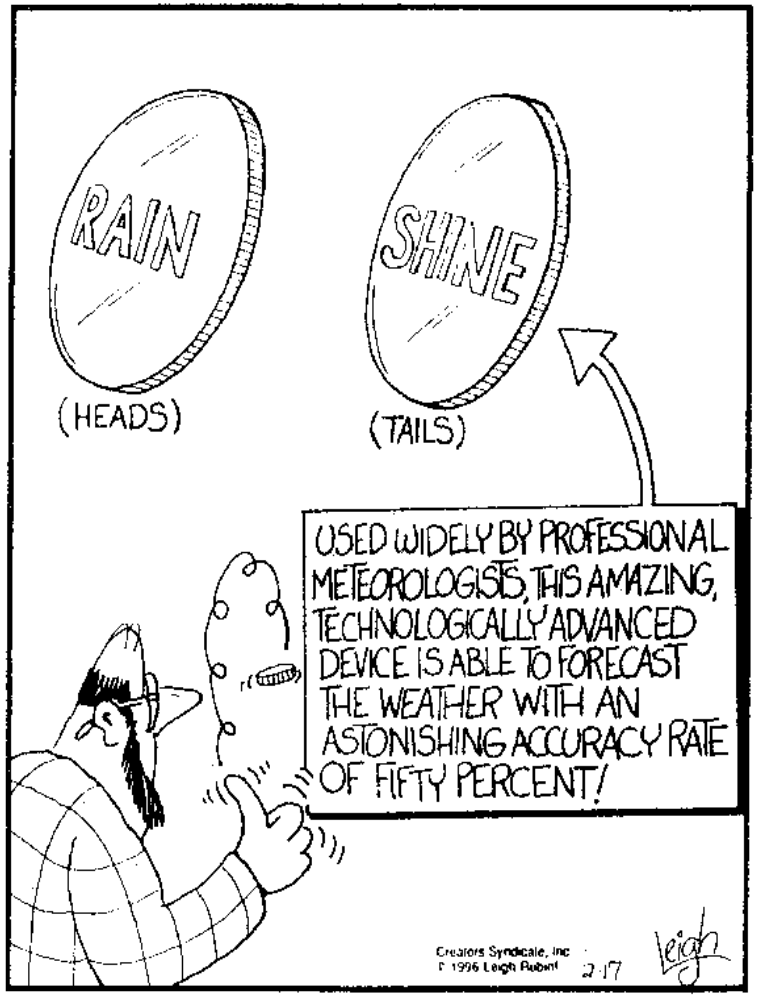


No warning/warning

X(t) = 0 or 1



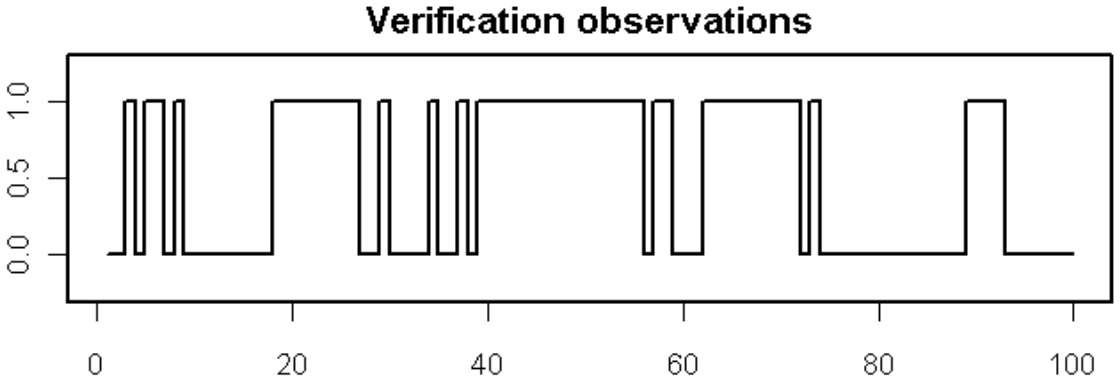
t = time (continuous/discrete)



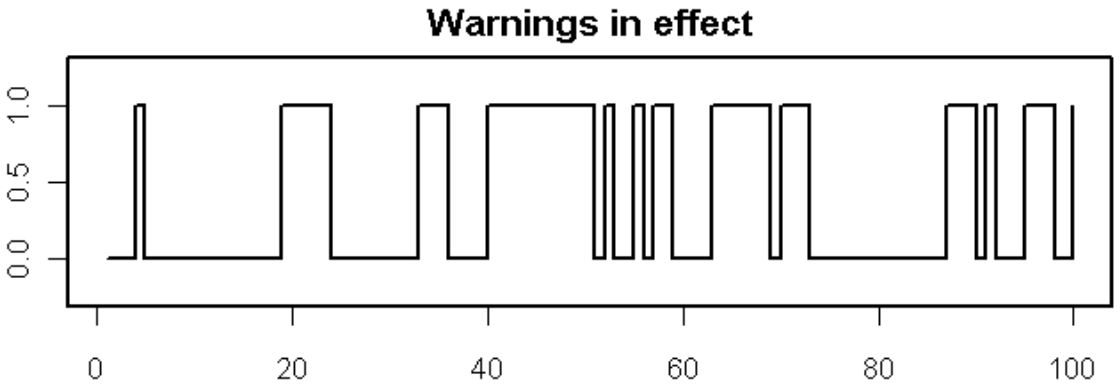
Some issues with observations:

- Sparseness of ground observations;
- Uncertainties in observations;
- Temporal coverage
- Spatial coverage within region of interest

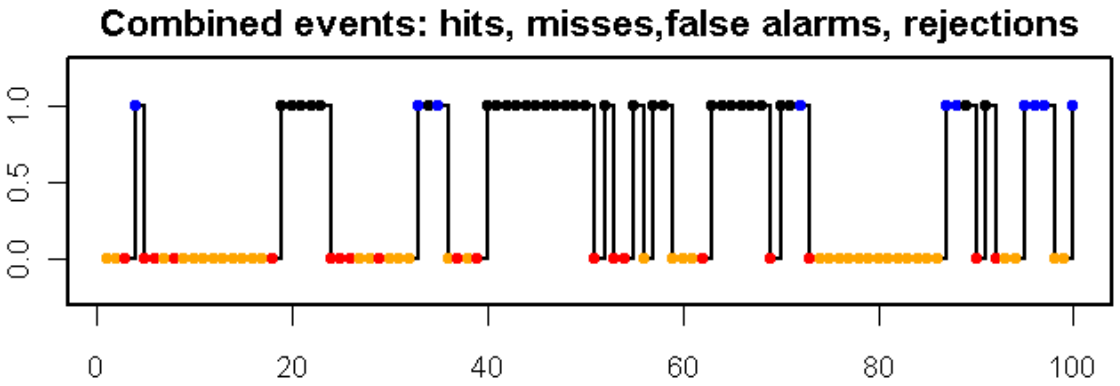
Artificial example: thresholded AR noise



$Y(t)$



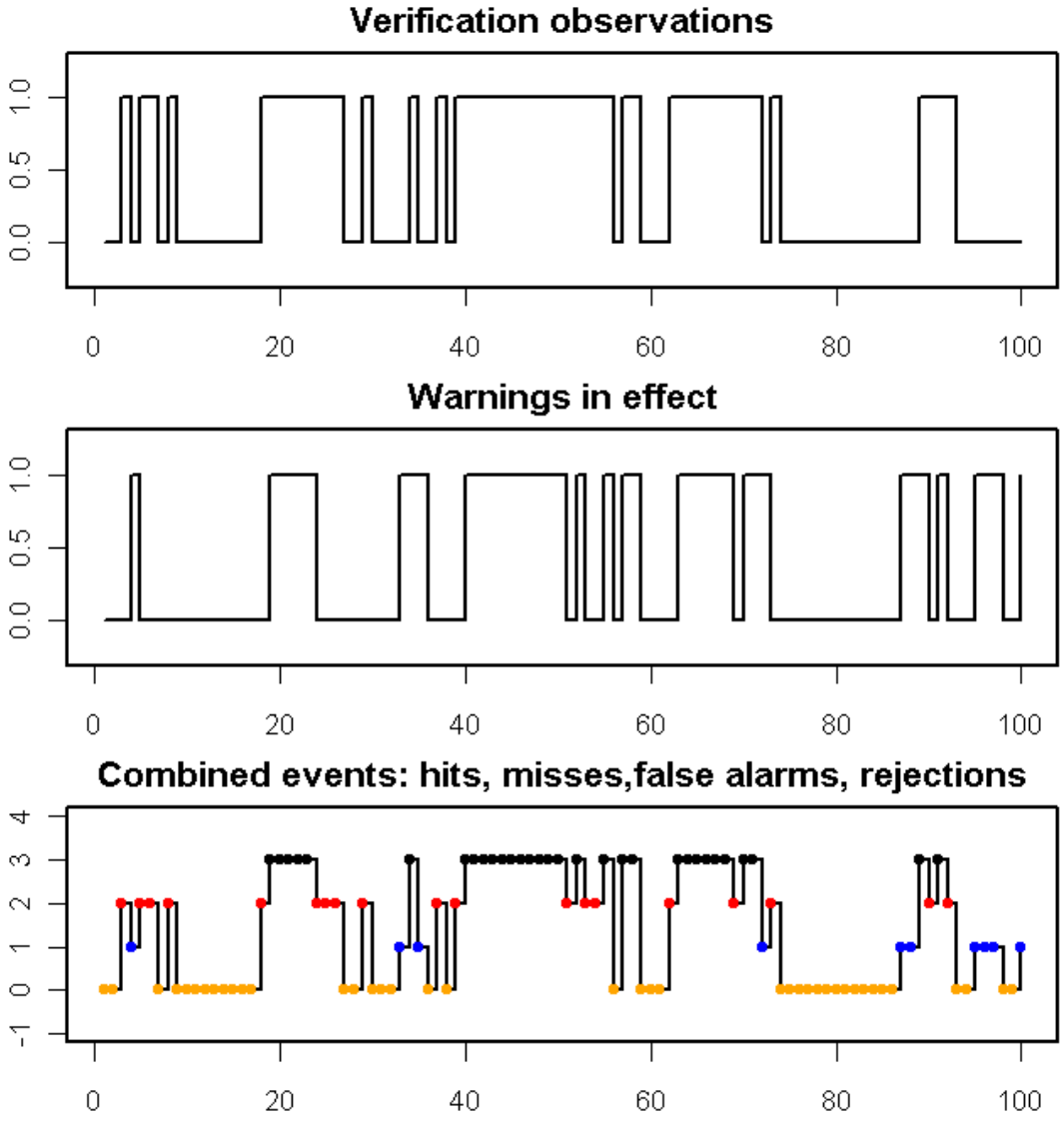
$X(t)$



Hit
Miss
False Alarm
Correct reject

How much skill is there?

Compound events



$Y(t)$

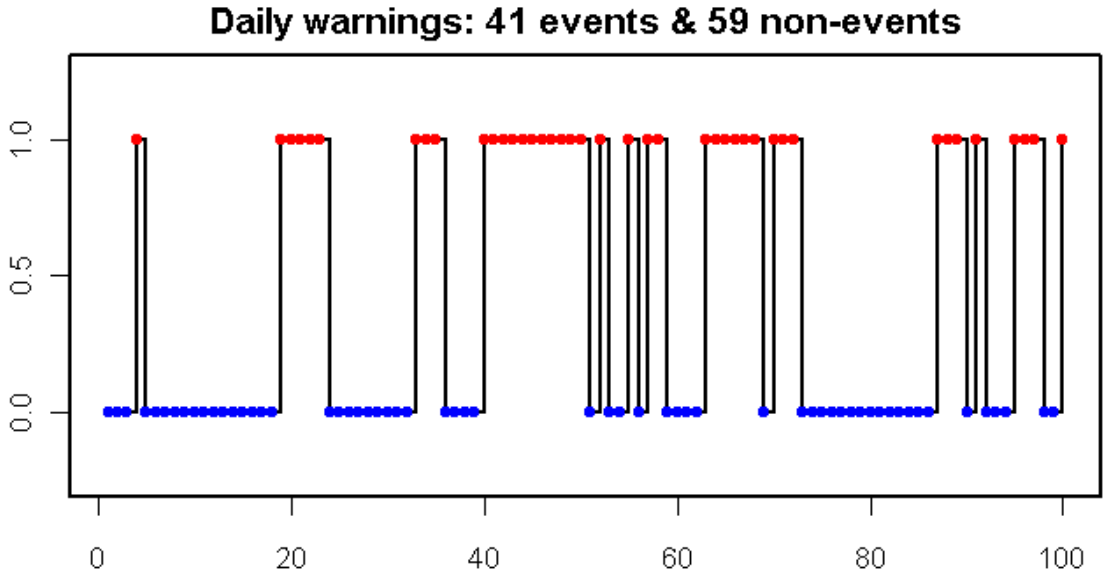
$X(t)$

$2Y + X$

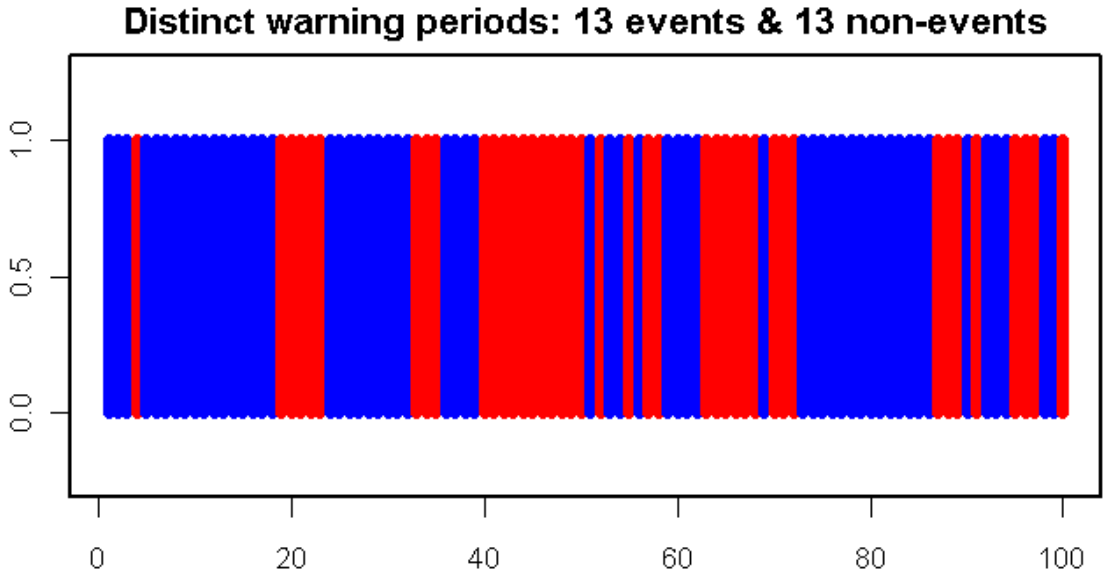
Hit
Miss
False Alarm
Correct reject

→ How to count the number of compound events?

How many events are there?



$$X(t)$$



→ Not obvious how to count events! (non-unique)

Contingency tables of counts

Daily events

	Obs Y=1	No-obs Y=0	
Warn X=1	31	10	41
No-warn X=0	19	40	59
	50	50	100

Distinct events

	Obs Y=1	No-obs Y=0	
X=1	10	7	17 (13)
X=0	15	12	27 (12)
	25 (12)	19 (12)	44

→ Get different counts depending upon how we count events (not only the number of correct rejections for rare events)

Yet another approach (thanks Chris Ferro)

Partition the time line into distinct warning periods and non-warning periods and then count the number of observed events in these periods.

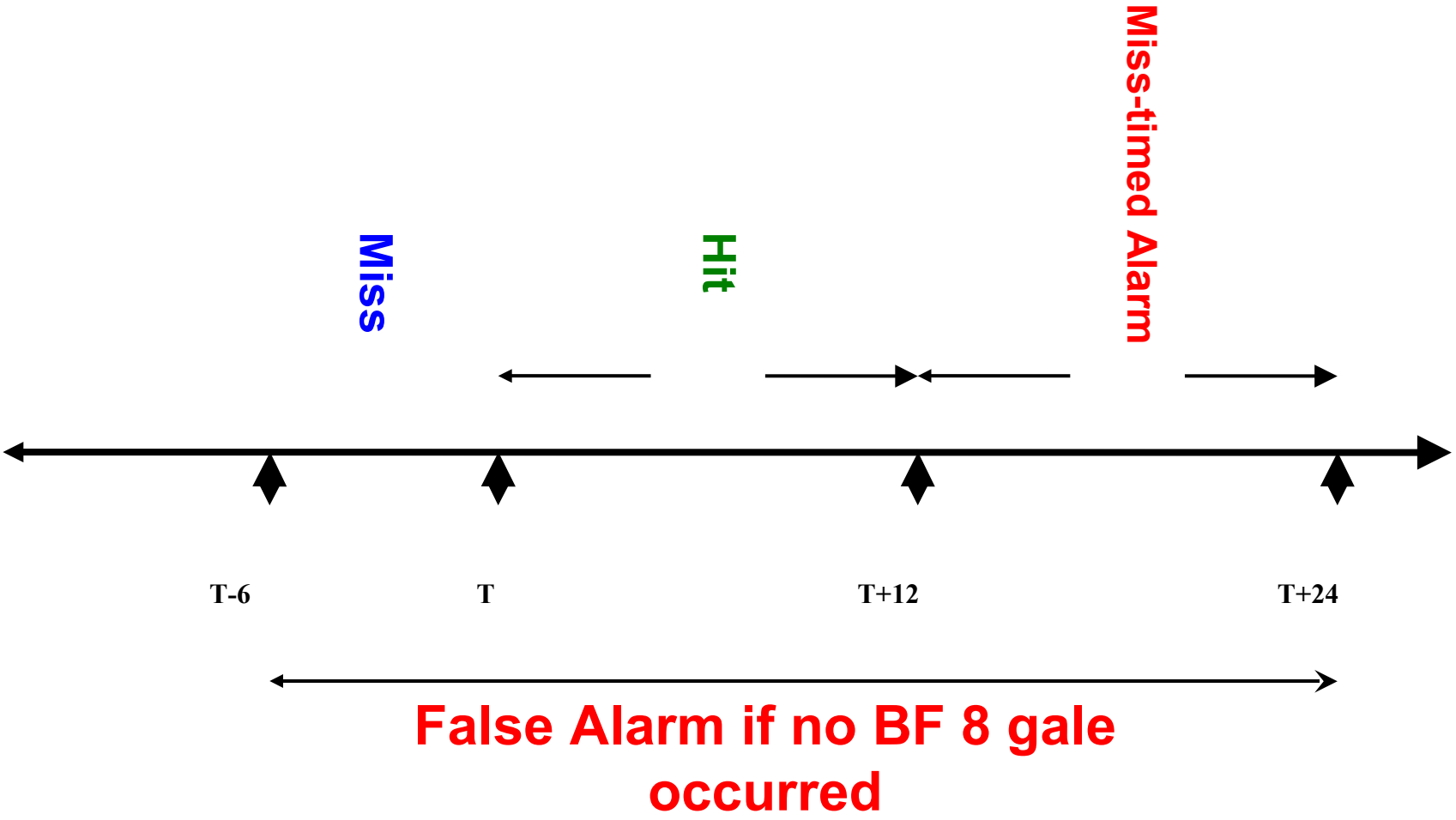
	Obs Y=1	No-obs Y=0	
X=1	19	19	38 (13)
X=0	18	18	36 (12)
	37 (12)	37 (12)	74

Obs during warnings	#Obs	Obs in non-warnings	#Obs
001	1	0010	1
01101000000001	3	1101000000000111111	3
111111111001000	2	111001000010	3
0100101	3	010111111111111	2
11111111111111	1	11	1
111	1	111	1
10	1	011	1
110001	2	000111111111	1
1111111	1	1110	1
1101000000000000000	2	10000000000000000001	2
0011	1	11	1
1100	1	100000	1
0	0	0	0
Total	**19**		**18**

Extended compound events

Near-misses, etc.

See Michael Sharpe's talk …



Note: no longer guaranteed to be a 2-way classification!

Verification measures

	Observed Y=1	No-obs Y=0	
Warning X=1	Hits a=npH	False Alarms b=np(B-H)	a+b=npB
No-warning X=0	Misses c=np(1-H)	Correct rejects d=n(1-p(1+B-H))	c+d=n(1-pB)
	a+c=np	b+d=n(1-p)	n=a+b+c+d

$$p = \frac{a+c}{n}$$ base rate

$$H = \frac{a}{a+c}$$ hit rate/POD

$$B = \frac{a+b}{a+c}$$ bias

$$C = 1 - FAR = \frac{a}{a+b} = \frac{H}{B}$$ "confidence"

$$\mathrm{T} = \frac{a}{a+b+c} = \frac{H}{1+B-H}$$ "threat score"/CSI

Comparison of verification measures

Measure	Daily counts	Distinct events	Partitioned events
Total counts n	100	43	74
Base rate p	0.50	0.57	0.50
Bias B	0.82	0.68	1.03
Hit rate H	0.62	0.40	0.51
Confidence C	0.76	0.59	0.50
Threat score T	0.52	0.31	0.34

→ Get different verification measures for different counting

→ Daily counting gives “best” skill values (H,C,T) for this example

→ BUT … we haven’t estimated sampling uncertainty on these measures!

Vanishing confidence & skill for rare events

$$C = 1 - FAR = \frac{a}{a+b} = \frac{H}{B} \quad \rightarrow \frac{\kappa p^{\delta}}{B} \rightarrow 0$$

$$\mathrm{T} = \frac{a}{a+b+c} = \frac{H}{1+B-H} \quad \rightarrow \frac{\kappa p^{\delta}}{1+B} \rightarrow 0$$

If bias stays finite, then skill measures such as confidence and threat score tend to zero for warnings of rarer events

→ Difficult to avoid "crying wolf" (issuing more false alarms than hits) for rare events (such as wolves!).

→ High confidence can only be maintained by reducing the bias (i.e. issuing fewer warnings) but then this will lead to more misses which is undesirable for many users.

Summary

- Although warnings are widely issued, it is not obvious how best to evaluate them (more research required!);

- Counting of compound events can be done in several different ways which all lead to different cell counts (and hence scores) – not just a problem with d! (number of correct rejections);

- Extended definitions of compound events are not always proper 2-way classifications and so interpretation of the 2x2 contingency table is no longer obvious;

- It is impossible to make high confidence (less false alarms than hits) unbiased warnings for rare events. ALL unbiased boys WILL generally cry wolf falsely on many occasions!

Ideas for future research

- Develop a more user-relevant way of evaluating warnings (i.e. move away from counting towards timings);

- Quantify the sampling uncertainty on the various scores (e.g. those for extended compound events). Note: dependency in time could make this challenging;

- Develop methods for spatial pooling of warnings of rare events that can deal with non-independence in space-time.

Some final words from Lao Tzu …

"Those who have knowledge, do not predict,
Those who predict, do not have knowledge."

- Lao Tzu

References

Stephenson, D.B., I.T. Jolliffe, C.A. Wilson, M. Sharpe, T.

Hewson, and M. Mittermaier, 2009:

White paper review on Weather Warnings,

Met Office Tech. Report (in revision).

Stephenson, D.B., Casati, B., C.A.T. Ferro, C.A. Wilson (2008):

The extreme dependency score:

a non-vanishing measure for forecasts of rare events

Meteorological Applications, Vol. 15, 41-50.

Our recent articles on verification

Casati, B., L. J. Wilson, D.B. Stephenson, P. Nurmi, A. Ghelli, M. Pocernich, U. Damrath, E. E. Ebert, B. G. Brown, S. Mason, (2008):

Forecast verification: current status and future directions,

Meteorological Applications, Special Issue: Forecast Verification, Vol. 15, 3-18.

Stephenson, D.B., Casati, B., C.A.T. Ferro, C.A. Wilson (2008):

The extreme dependency score: a non-vanishing measure for forecasts of rare events,

Meteorological Applications, Special Issue: Forecast Verification, Vol. 15, 41-50.

Doblas-Reyes, F.J., C. A. S. Coelho, D. B. Stephenson (2008):

How much does simplification of probability forecasts reduce forecast quality?,

Meteorological Applications, Special Issue: Forecast Verification, Vol. 15, 155-162.

Jolliffe, I.T. and Stephenson, D.B. (2008):

Proper Scores for Probability Forecasts Can Never Be Equitable,

Monthly Weather Review, Vol. 136, No. 4., 1505–1510.

Verification papers 2003-2007

Jolliffe, I.T. (2007)
Uncertainty and inference for verification measures,
Weather and Forecasting, 22, pp 137-150

Mailier, P.J., I.T. Jolliffe, and D.B. Stephenson, (2006)
Quality of Weather Forecasts: Review and Recommendations,
Royal Meteorological Society Project Report, pp. 89.

Jolliffe, I.T. and D.B. Stephenson (2005)
Comments on "Discussion of Verification Concepts in Forecast Verification: A Practitioner's Guide in Atmospheric Science",
Weather and Forecasting, 20, pp 796-800

Göber, M., C.A. Wilson, S.F. Milton, D.B. Stephenson (2004):
Fairplay in the verification of operational quantitative precipitation forecasts,
J. Hydrology, 288, pp 225-236

Casati, B., G. Ross, and D.B. Stephenson (2004):
A new intensity-scale approach for the verification of spatial precipitation forecasts,
Meteorological Applications, 11, pp 141-154

Downloadable as pdf files from:
http://www.secam.ex.ac.uk/xcs

Forecast verification

is the exploration and assessment of the quality of a forecast system inferred from a sample of pairs of previous forecasts and observations:

Good forecasts have:

- Utility (value)
- Consistency
- Reliability (i.e. unbiased: E(Y|X)=X)
- Resolution (i.e. reduce uncertainty: var(Y|X) < var(Y))
- Accuracy
- Skill/Association

…

→ More than 1 score needed to summarize forecasts.

A. H. Murphy 1993
"What is a good forecast ?
An essay on the nature of goodness in weather forecasting"
Weather and Forecasting, 8, 281-293.