Verification of warnings and extremes
-
issues and approaches

Martin Göber

Hans-Ertel-Centre for Weather Research (HErZ)
Deutscher Wetterdienst DWD
E-mail: martin.goeber@dwd.de

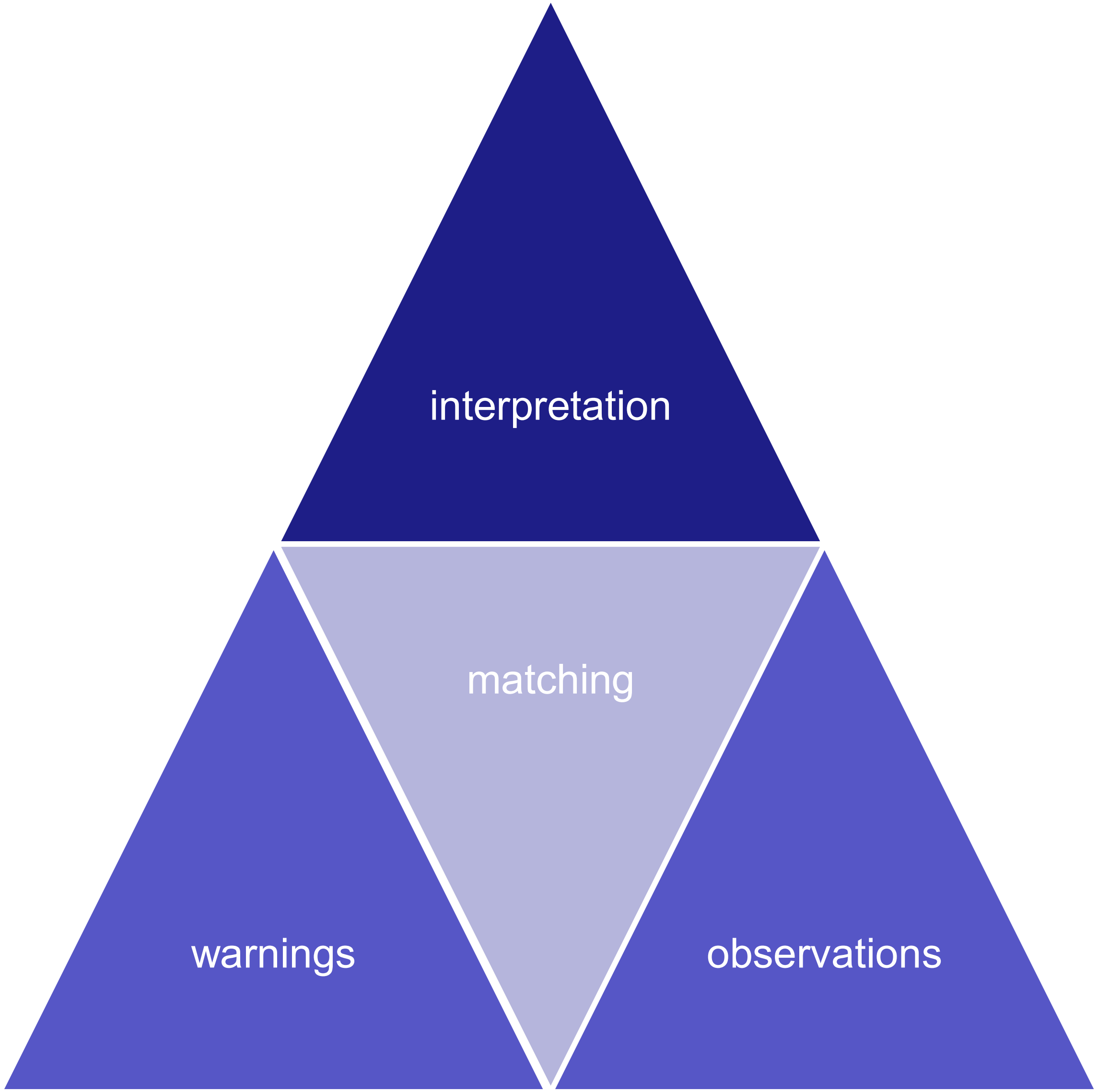
Summary

Users of warnings are very diverse and thus warning verification is also very diverse.

Each choice of a parameter of the verification method has to be user oriented – there is no „one size fits all“.

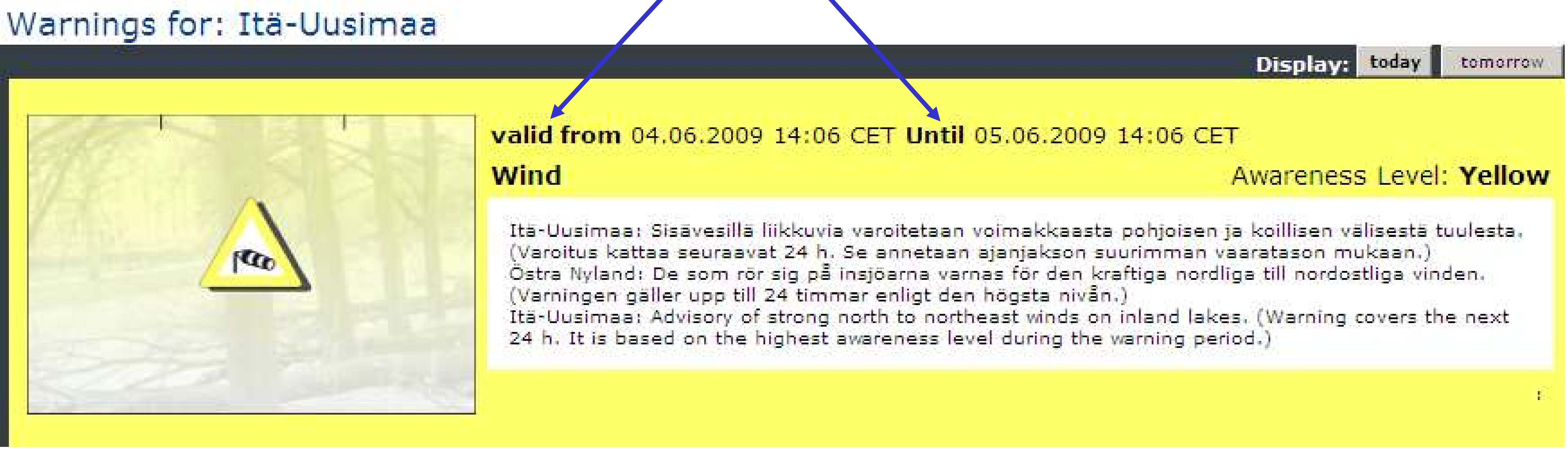



Outline



Warnings

2 additional free parameters
when to start: **lead time**
how long: **duration**



These additional free parameters have to be decided upon by:

- the forecaster, or
- fixed by process management (driven by user needs)

Issue: physical thresholds

Warnings:

- clearly defined thresholds/events, yet some confusion since either as country-wide definitions or adapted towards the regional climatology
- sometimes multicategory (“winter weather”, “thunderstorm with violent storm gusts”, “thunderstorm with intense precipitation”)
- **worst thing** possible in an area, or worst thing in a “significant” part of the area

Observations:

- clearly defined at first glance
  - yet warnings are mostly for areas, events localised → undersampling
  - “soft touch” required because of overestimate of false alarms
    - use of “practically perfect forecast” (Brooks et al. 1998)
    - allow for some overestimate, since user might be gracious, as long as something serious happens
    - ultimately: probabilistic analysis of events needed

Issue: physical thresholds

gust warning verification, winter

”one category too high, is still ok, →no false alarm”

“severe”

	observed gusts in m/s or Bft							absolute frequ.	FAR	soft FAR	difference
	<14	14-17	18-24	25-28	29-32	33-37	>38				
	0-6	7	8-9	10	11	12	>12				
warnings											
no warning	**561834**	5244	300	1	0	0	0	567379			
near gale	66927	**19312**	1810	10	0	0	0	88059	0,59		
gale	23850	22227	**11036**	262	21	1	0	57397	0,75	0,37	0,37
storm	1295	2231	3557	**391**	52	3	0	7529	0,91	0,44	0,47
violent storm	207	577	1052	251	**80**	11	2	2180	0,96	0,84	0,12
hurrican force	136	208	414	118	37	**7**	1	921	0,99	0,95	0,04
extreme hurrican f.	0	0	0	0	0	0	**0**	0			
absolute frequency	654249	49799	18169	1033	190	22	3	723465			

“severe”

Issue: observations

<u>What:</u>

- standard: SYNOPS
- increasingly: lightning (nice! :), radar
- non-NMS networks
- "citizen observations – posting about the weather and it's impacts":
    - dedicated mobile apps, social media (twitter, Instagram photo descriptions), spotters(e.g. European Severe Weather Database ESWD)

<u>Data quality:</u>

- particularly important for warning verification
- "skewed verification loss function": missing to observe an event is not as bad as falsely reporting one and thus have a *missed* warning
- multivariate approach strongly recommended (e.g. severe rain in synop wrong, where there was no radar or satellite signature)

Issue: matching warning and obs

Largest difference to model verification !

temporal

- hourly (SYNOPS), e.g. NCEP, UKMO, DWD as “process oriented verification”
- “events”:
  - warning and/or obs immediately followed by warning
  - obs in an interval starting at first threshold exceedance (e.g. UKMO 6 hours before the next event starts)
  - even “softer” definition: as “extreme events”
- thus size of sample N varies between a few dozens and millions !
- lead time for a hit: desired versus real; 0, 1, … hours ?

Met Office warning ver

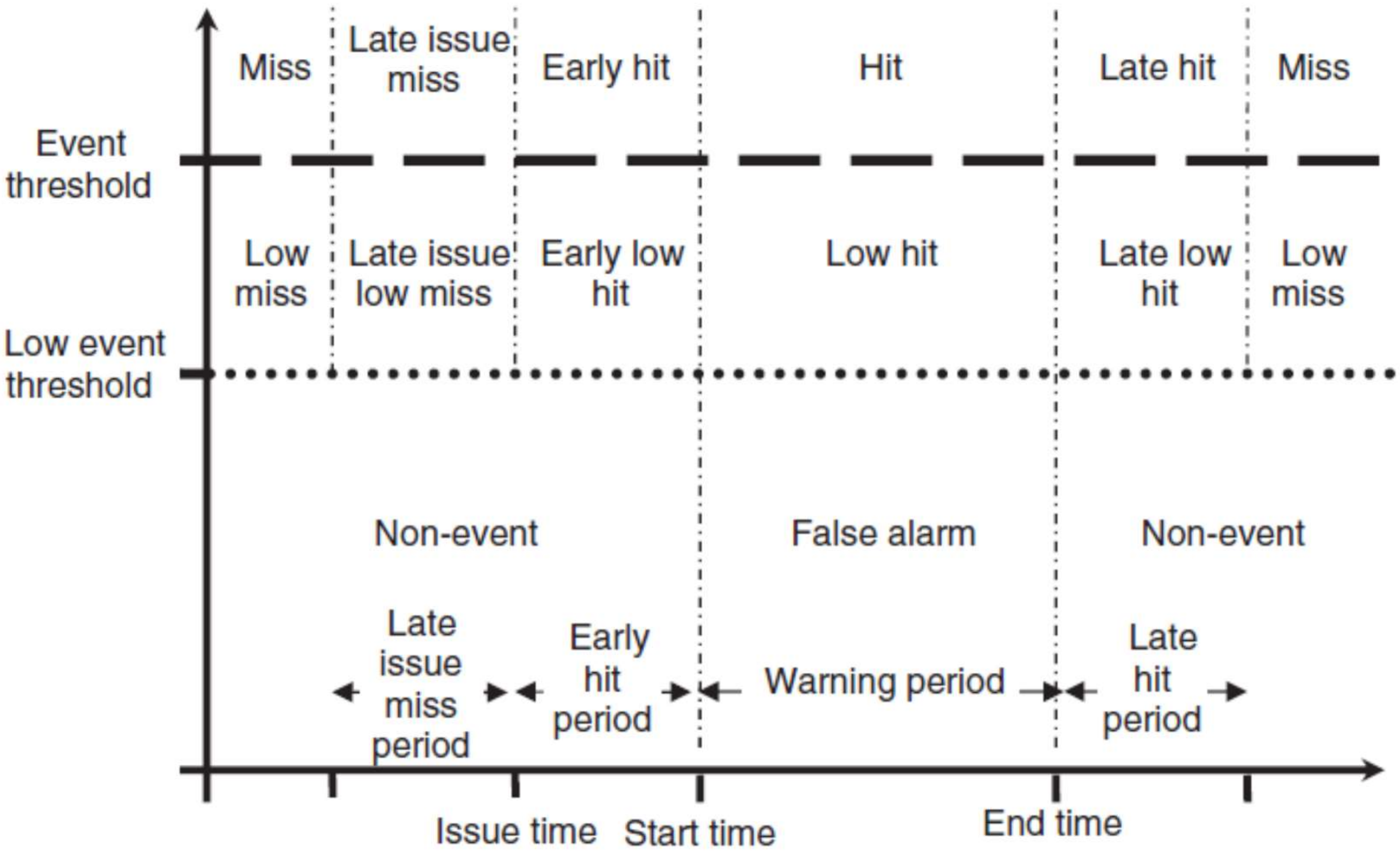


Figure 3. Event categories for flexible verification.

Sharpe, M. (2016): A flexible approach to the objective verification of warnings. Met. Applications

Issue: matching warning and obs

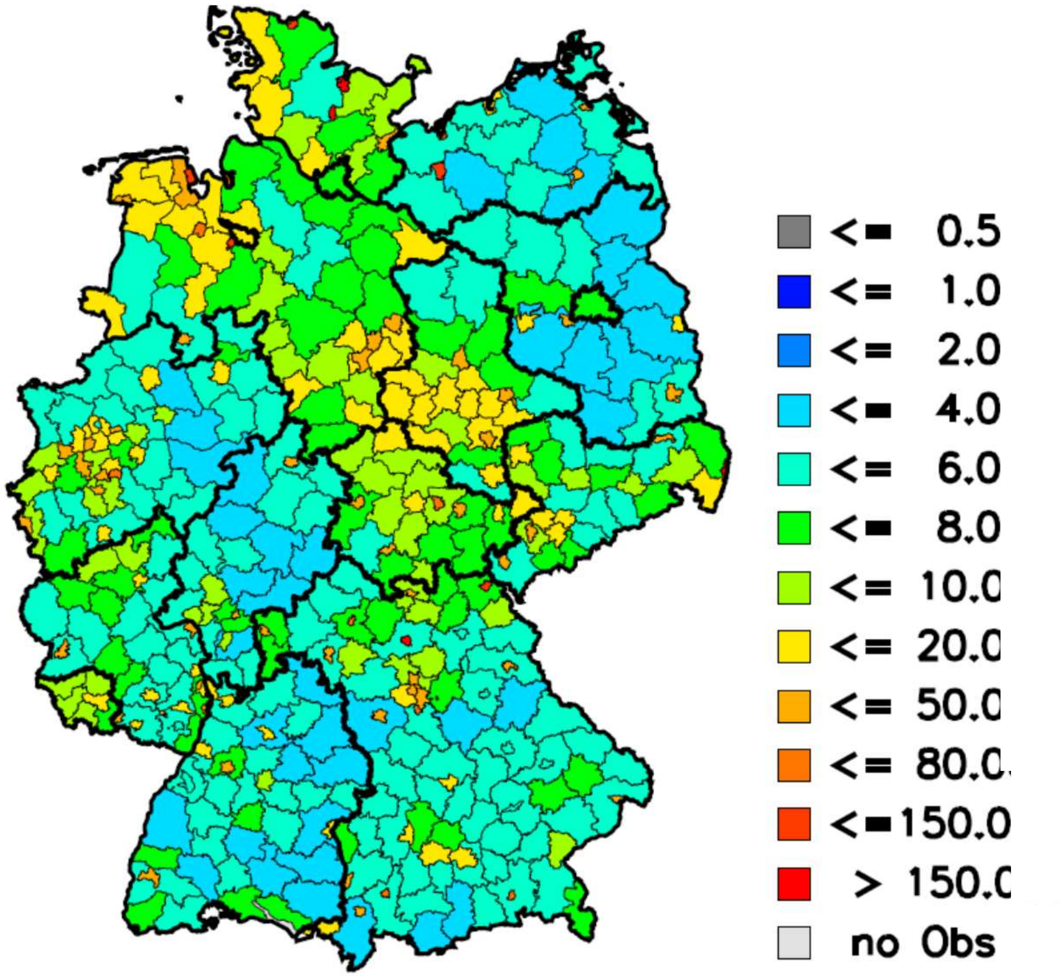
Largest difference to model verification !

spatial

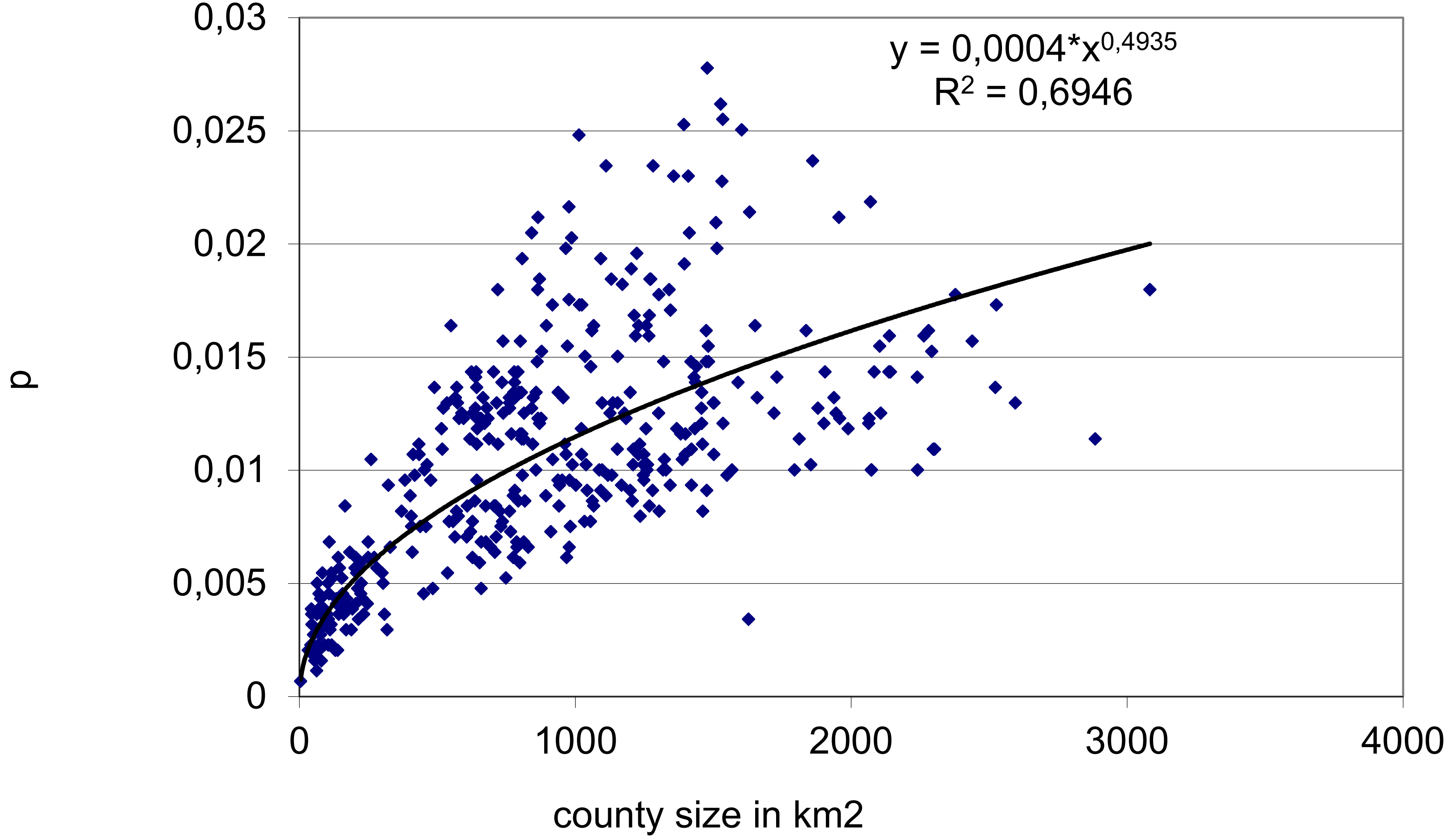
- sometimes “by-hand” (e.g. Switzerland, France)
- worst thing in the area
- “MODE-type” (**M**ethod for **O**bject-based **D**iagnostic **E**valuation)
- dependency on area size possible
    - example: thunderstorm warning ver against lightning obs (continuous in space and time!)

Issue: matching warning and obs

Thunderstorms (lightning): frequency bias

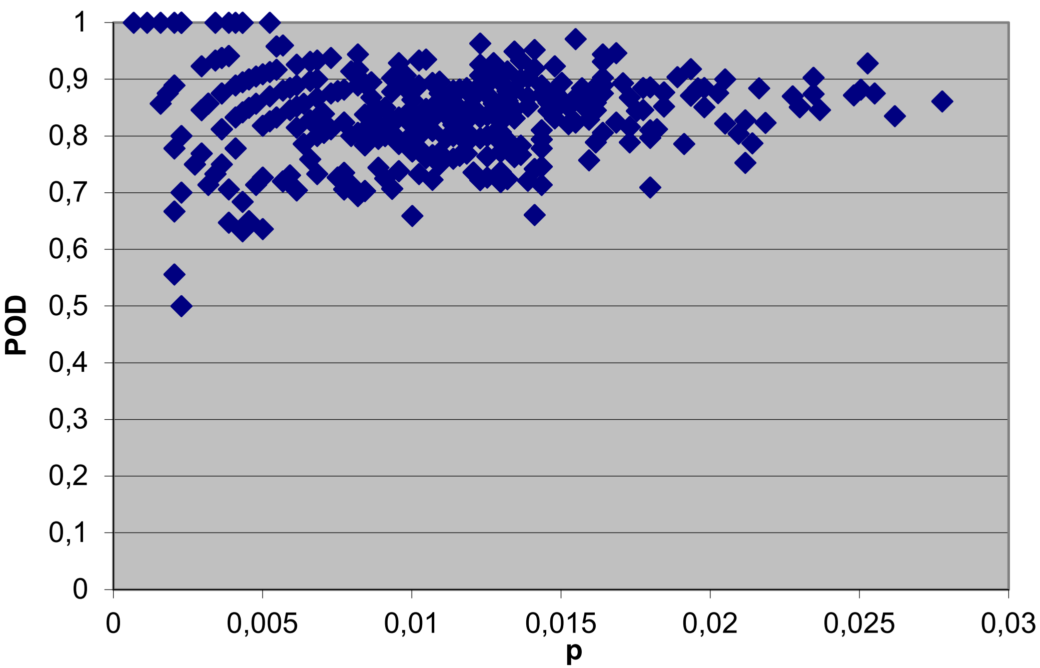


Base rate

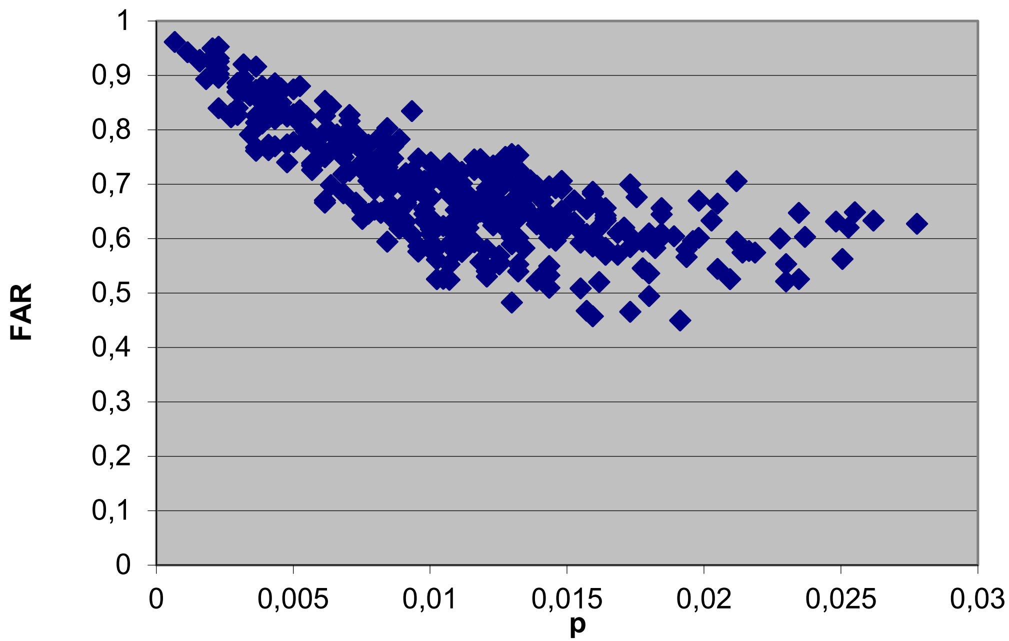


y = 0,0004*x^0,4935
R² = 0,6946
0,03
0,025
0,02
0,015
0,01
0,005
0
p
0
1000
2000
3000
4000
county size in km2

POD



FAR



p = base rate in thunderstorms / hour

Issue: measures

- “everything” used (including extreme dependency scores, ROC-area)
- POD (view of the media: “something happened, has the weather service done it’s job ?”)
- FAR (view of an emergency manager: “the weather service activated us, was it justified ?”
- threat score (or “Critical Success Index” CSI) frequently used, since definition of the no-forecast/no-obs category sometimes seen as problematic
    - yet CSI can be easily hedged by overforecasting
    - way out: no-forecast/no-obs category can be defined by using regular intervals of no/no (e.g. 3 hours) and count how often they occur

Issue: measures

Finley dataset, 1884

Tornado forecast	Tornado observed		
	Yes	No	fc Σ
Yes	28	72	100
No	23	2680	2703
obs Σ	51	2752	2803

Percent correct:

Finley: 97%
Never tornado: 98 %

Beware of score behaviour for rare (interesting) events

Slide from Laurie Wilson's talk on categorical ver.

- **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

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

Base rate independence ⇔ Functions of H and F

Extremal Dependency Index - EDI

Symmetric Extremal Dependency Index - SEDI

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

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

Issue: measures

Wilson, L., Giles, A. (2013): A new index for the verification of accuracy and timeliness of weather warnings . Met. Applications

For one variable:

$$WWI = AS + 0.5 * \frac{LTR - 1}{LTR_{\max} - 1}(1 - AS)$$

$$LTR = \frac{LT}{TLT}$$

$$AS = EDI = \frac{\ln(F) - \ln(H)}{\ln(F) + \ln(H)}$$

WWI: Weather Warning Index

LT: (average) lead time
TLT: Target Lead Time
LTR: Lead Time Ratio
LTR_{max}: max. benefit for long lead

AS: accuracy score

Issue: “Interpretation” of results

Performance targets:

• extreme interannual variability for extreme events

• strong influence of change of observational network; “if you detect more, it’s easier to forecast” (e.g. apparently strong increase in skill after NEXRAD introduction in the USA)

Case studies

• remain very popular, rightly so ?

Significance

• only bad if you think in terms wanting to *infer* future performance, ok if you just think *descriptive* about what has happened

• care needed when extrapolating from results for mildy severe events to very severe ones, since there can be step changes in forecaster behaviour taking some Cost/Loss ratio into account

Issue: “Interpretation” of results

<u>Consequences</u>

- changing forecasting process
  - e.g shortening of warnings at DWD dramatically reduced false alarm ratio based on hourly verification almost without reduction in POD
  - in the USA, move from county based to polygon based warnings strongly reduced spatial overforecasting
  - creating new products (probabilistic forecasts)

Issue: user-based assessments

- important role, especially during process of setting up county based warnings and subsequent fine tuning of products, given the current ability to predict severe events
- surveys, user workshops, direct observations, public opinion monitoring, feedback mechanisms, anecdotal information
- presentation of warnings to the users essential
- “vigilance evaluation committee” (Meteo France /Civil Authorities), SWFDP in Southern Africa, MAP-D-Phase
- typical questions:
    - Do you keep informed about severe weather warnings?
    - By which means?
    - Do you know the warning web page and the meaning of colours?
    - Do you prefer an earlier, less precise warning or a late, but more precise warning?
    - ……………

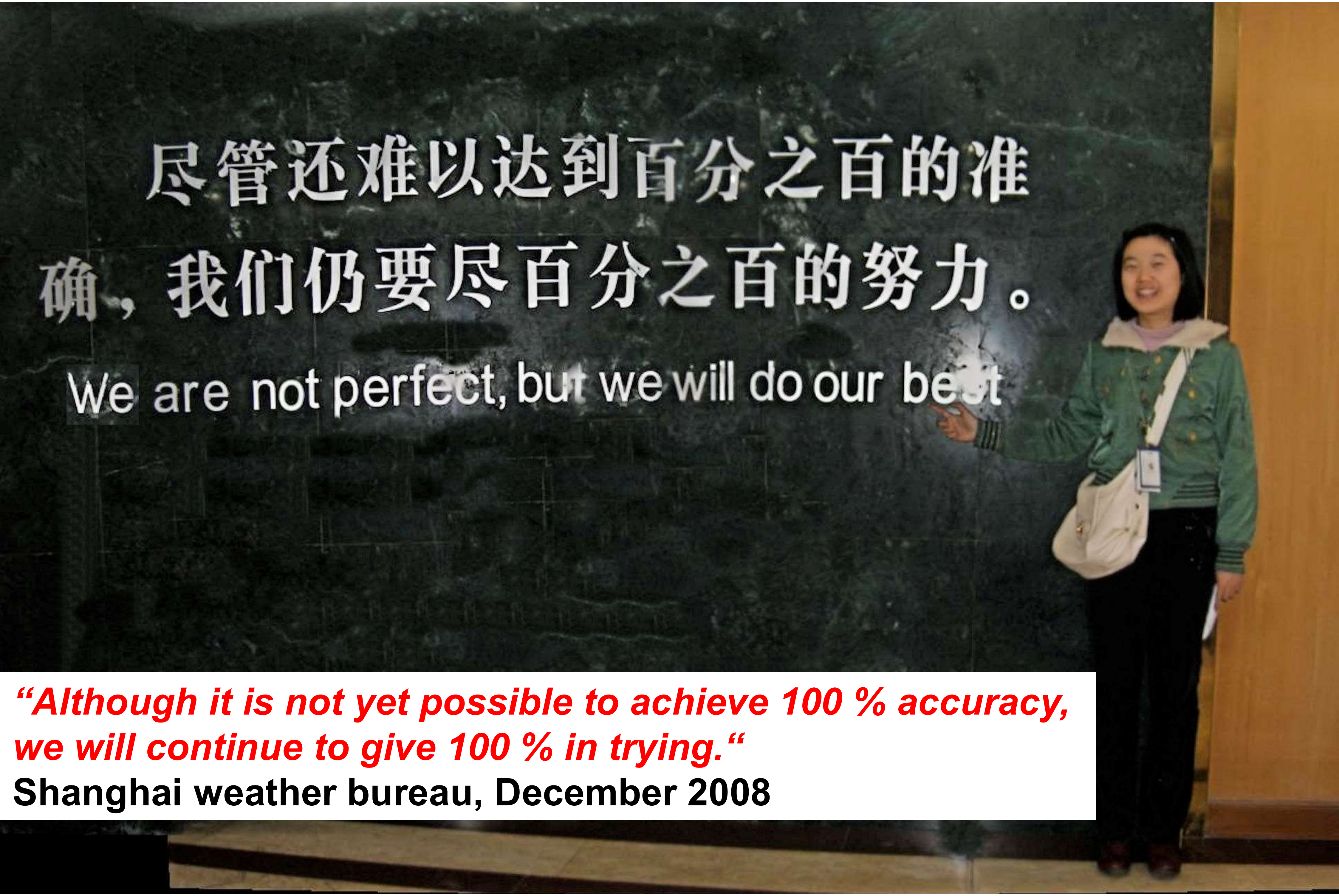
Issue: Comparing warning guidances and warnings

- End user verification: verify at face value
- Model (guidance) verification: measure potential

Summary

Users of warnings are very diverse and thus warning verification is also very diverse.

Each choice of a parameter of the verification method has to be user oriented – there is no „one size fits all“.



"Although it is not yet possible to achieve 100 % accuracy, we will continue to give 100 % in trying."

Shanghai weather bureau, December 2008