



Warning verification - definition and approaches

Martin Göber

Department Weather Forecasting
Deutscher Wetterdienst DWD
E-mail: martin.goeber@dwd.de



Summary

Users of warning are either direct and handling warning verification is also either direct.

Each choice of a parameter of the verification method has to be well oriented here in no one is efficient.



Disposition Q & A (pproaches)

1. Information about warning verification (5)
2. Characteristics of warnings (10 minutes)
3. Observations: which, sparseness, quality, thresholds (10)
4. Matching of warnings and observations (15)
5. Measures (10)
6. interpretation of results, user based assessments (20)
7. Comparison of warning guidances with warnings (15)



Issue: state of available information

19 out of 26 students answered (at least 1 question)
= 73 % answer rate

3. Are there documents in your service which lay down the rules in warning verification (or which generally describe how warning verification is done)?			Response Percent	Response Count
Yes			33.3%	6
No			44.4%	8
don't know			22.2%	4
			<i>answered question</i>	18
			<i>skipped question</i>	1



Issue: state of available information

- Warning verification is hardly touched in the „bibles“, i.e.: Wilks statistics textbook; Jolliffe/Stephenson’s verification book; Nurmi’s ECMWF „Recommandations on verification of local forecasts“; THE JWGV web-page, some mentioning in Mason’s consultancy report.
- Yet lots of the categorical statistics can be used, although additional care is needed.
- It’s very difficult to find information on the web or otherwise about the NMS’ procedures – exception: NCEP’s hurrican and tornado warnings.
- What is clear: compared to model verification it is **rpri ingl di er e**, because it should be (often is) **dri en b di er e er** .
- One document has quite a lot of information concentrated on user-based assessments: WMO/TD No. 1023 *Guidlines on performance assessment of public weather services*. (Gordon, Shaykewich, 2000).
<http://www.wmo.int/pages/prog/amp/pwsp/pdf/TD-1023.pdf>



Information sources

Presentation based on (partly scetchy) information from NMS of 10 countries (Thanks!):

- Austria
- Botswana
- Denmark
- Finland
- France
- Germany
- Greece
- Switzerland
- UK
- USA



Warnings

European examples of warnings

<http://www.meteoalarm.eu>



meteoalarm
 alerting europe for extreme weather

<http://www.meteoalarm.eu>

[Start](#) | [News](#) | [About Meteoalarm](#) | [Help](#) | [Terms and Conditions](#) | [Links](#) | [Greyscale Maps](#)

» Europe:

Weather warnings: Europe:

awareness types:
 Display:

Created: 04.06.2009 14:51:06 | Valid for: 04.06.2009

Awareness Reports
 You can find detailed information about relevant country.

AT					
BE					
CH					
CY					
CZ					
DE					
DK					
ES					
FI					
FR					
GR					
HU					
IE					
IS					
IT					
LU					
MT					
NL					
NO					
PL					
PT					
RO					
SE					
SI					
SK					
UK					

Legend:
 White: No awareness
 Green: Low awareness
 Yellow: Moderate awareness
 Orange: High awareness
 Red: Extreme awareness
 Wind: Wind
 Rain: Rain
 Snow/Ice: Snow/Ice
 Thunderstorms: Thunderstorms
 Fog: Fog
 Extreme high temperature: Extreme high temperature
 Extreme low temperature: Extreme low temperature
 Coastal Event: Coastal Event
 Forestfire: Forestfire
 Avalanches: Avalanches



Warnings

Yellow:

1. The weather is potentially dangerous. The weather phenomena that have been forecast are not unusual,
2. but be attentive if you intend to practice activities exposed to meteorological risks.
3. Keep informed about the expected meteorological conditions and do not take any avoidable risk.

Orange:

1. The weather is dangerous. Unusual meteorological phenomena have been forecast.
2. Damage and casualties are likely to happen.
3. Be very vigilant and keep regularly informed about the detailed expected meteorological conditions. Be aware of the risks that might be unavoidable. Follow any advice given by your authorities.

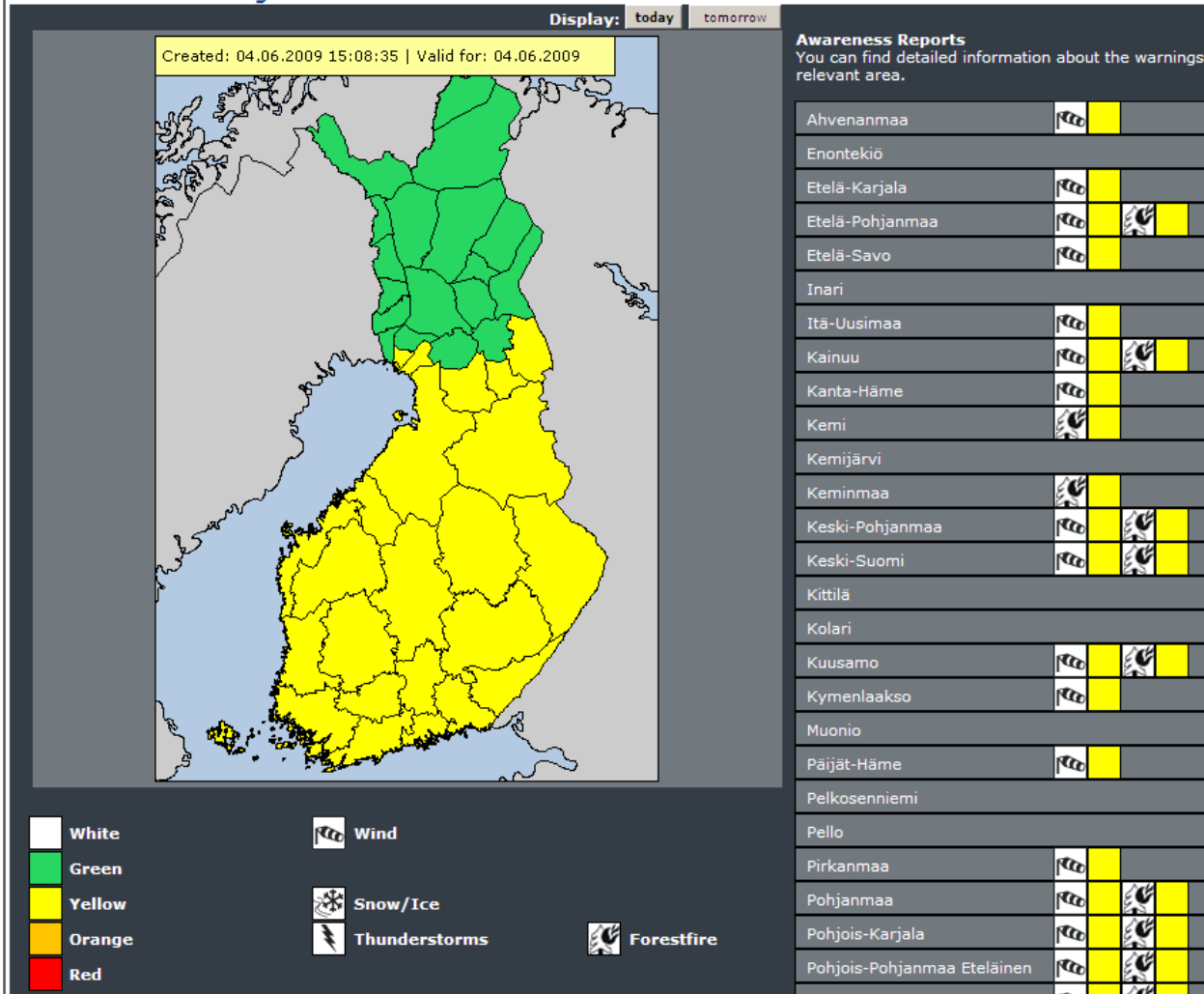
Red:

1. The weather is very dangerous. Exceptionally intense meteorological phenomena have been forecast.
2. Major damage and accidents are likely, in many cases with threat to life and limb, over a wide area.
3. Keep frequently informed about detailed expected meteorological conditions and risks. Follow orders and any advice given by your authorities under all circumstances, be prepared for extraordinary measures.



Warnings

Weather warnings: Finland



Paradigm shift in 21st ct:
many warnings issued on
a small, regional scale



Warnings

1. What are the spatial scales on which warnings are issued and/or verified in your country? Check all that apply.

	issued	verified	verification rate	Response Count
cities	100.0% (10)	60.0% (6)	60 %	10
counties	100.0% (4)	50.0% (2)	50 %	4
provinces	100.0% (12)	58.3% (7)	58 %	12
whole country	88.9% (8)	77.8% (7)	88 %	9
	<i>answered question</i>			19
	<i>skipped question</i>			0



Warnings

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

Warnings for: Itä-Uusimaa

Display: today tomorrow

valid from 04.06.2009 14:06 CET **Until** 05.06.2009 14:06 CET

Wind Awareness Level: **Yellow**

Itä-Uusimaa: Sisävesillä liikkuvia varoitetaan voimakkaasta pohjoisen ja koillisen välisestä tuulesta. (Varoitus kattaa seuraavat 24 h. Se annetaan ajanjakson suurimman vaaratason mukaan.)
Östra Nyland: De som rör sig på insjöarna varnas för den kraftiga nordliga till nordostliga vinden. (Varningen gäller upp till 24 timmar enligt den högsta nivån.)
Itä-Uusimaa: Advisory of strong north to northeast winds on inland lakes. (Warning covers the next 24 h. It is based on the highest awareness level during the warning period.)

These additional free parameters have to be decided upon by the forecaster
or
fixed by process management (driven user needs)



Warnings

2. What is the required lead time for warnings for areas of different size? Check all that apply.

	cities	counties	provinces	whole country	Response Count
don't know	0.0% (0)	0.0% (0)	0.0% (0)	0.0% (0)	0
less than 1 hour	100.0% (2)	50.0% (1)	0.0% (0)	0.0% (0)	2
1 hour	50.0% (1)	50.0% (1)	0.0% (0)	0.0% (0)	2
2 hours	100.0% (2)	0.0% (0)	0.0% (0)	0.0% (0)	2
3 hours	0.0% (0)	0.0% (0)	0.0% (0)	100.0% (1)	1
6 hours	33.3% (1)	66.7% (2)	33.3% (1)	33.3% (1)	3
half a day	50.0% (2)	25.0% (1)	25.0% (1)	50.0% (2)	4
1 day	22.2% (2)	33.3% (3)	66.7% (6)	55.6% (5)	9
2 days	12.5% (1)	37.5% (3)	62.5% (5)	87.5% (7)	8
don't know	40.0% (2)	40.0% (2)	60.0% (3)	40.0% (2)	5
<i>answered question</i>					18

skipped question 1

grey highlighting: highest value in each row%
endenc : larger scale, larger lead time



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)

Observations:

- clearly defined at first glance
 - yet warnings are mostly for areas → 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
 - probabilistic analysis of events needed



Issue: physical thresholds

gust warning verification, winter

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

"severe"

"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
hurricane force	136	208	414	118	37	7	1	921	0,99	0,95	0,04
extreme hurricane f.	0	0	0	0	0	0	0	0			
absolute frequency	654249	49799	18169	1033	190	22	3	723465			



Issue: observations

4. What kind of observations do you use to verify warnings? Check all that apply.

	Response Percent	Response Count
synoptical observations	93.3%	14
metars	46.7%	7
non-meteorological networks	13.3%	2
lightning	20.0%	3
radar	33.3%	5
satellite	53.3%	8
media reports of damages	46.7%	7
spotters	6.7%	1
other eye witness reports	26.7%	4
Hide replies Other (please specify)		4

1. AIREPS Fri, Jun 5, 2009 9:42 AM [Find...](#)
2. I don't know how they verify warnings (is a forecaster's task and I'm working on numerical models). I think they don't have any formal procedure to verificate warnings. Thu, Jun 4, 2009 9:57 PM [Find...](#)
3. PM10 for yellow dust
AWS Thu, Jun 4, 2009 5:01 PM [Find...](#)
4. Agriculture and Hidrology reports Thu, Jun 4, 2009 12:25 PM [Find...](#)

	<i>answered question</i>	15
	<i>skipped question</i>	4



Issue: observations

5. How do you deal with the sparseness of the observations, i.e. how many observations (in the area and/or time interval) do you require to have been above a threshold for saying an "event has occurred"? How do you deal with the quality of the observations? Check all that apply.

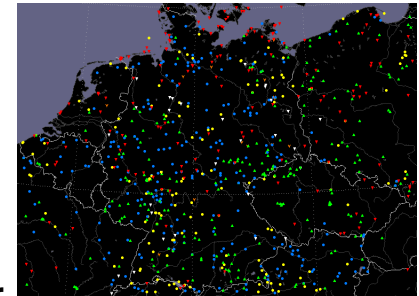
	yes	no	Response Count
single obs above threshold sufficient	50.0% (6)	50.0% (6)	12
single obs slightly below threshold already sufficient	33.3% (3)	66.7% (6)	9
more than one obs needed	86.7% (13)	13.3% (2)	15
data are automatically quality controlled	40.0% (4)	60.0% (6)	10
data are manually quality controlled	83.3% (10)	16.7% (2)	12
	<i>answered question</i>		16
	<i>skipped question</i>		3



Issue: observations

What:

- standard: SYNOPS
- increasingly: lightning (nice! :), radar
- non-NMS networks
- additional obs from spotters, e.g. European Severe Weather Database ESWD



Data quality:

- 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. no severe precip, where there was no radar or satellite signature)



Issue: data formats

Warnings:

- all sorts of ASCII formats, yet trend towards xml

Observations:

- "standard chaos"
- additional obs from spotters, ASCII, ESWD

Raw ASCII data (Selected event: F4 Pforzheim tornado, Germany, 10 July 1968)

INFO|10|V01.40|3|QC2|EYEWITN LIT NWSP TV WXSVC WWW|TorDACH V1.6.00, tordach.org/de, de@tordach.org; D. Fuchs, Promet 4'81, 8-10 ==> Monatl. Witterungsber. DWD;; Monatsarbeit der Wetterdienst-Referendarausbildung, 1978, 56 S.;; Pers. comm. 2000; R. Nestle, Meteor. Rdschau. 22 (1969), 1-3; Becht H. P., Stadtarchiv Pforzheim, pers. comm. (1998); Fulks, H.W., 1969: A synoptic review of the Pforzheim tornado of; 10 July 1968. 2nd Wea. Wing Tech. Bull, Air Wea. Service, US Air Force;; April 1969, 26-43.|1|Nikolai Dotzek, ESSL|20051231

TIME&PLACE|19|1968|07|10|WED|20|30|1H|DE|BW|Ittersbach, Pforzheim||48.9055|08.5270|HILLS|RURAL|RURAL URBAN

TORNADO|23||4|8|DMGTEXT||FNLOBS|NOSVTCOBS|||20|35||1000|W-E|DM:150M|170kFm||300|2|Same cell as T7/F3 tornado at Sarrebourg, Eschbourg, Hagenau#



Issue: matching warning and obs

Large difference of model verification !

temporal

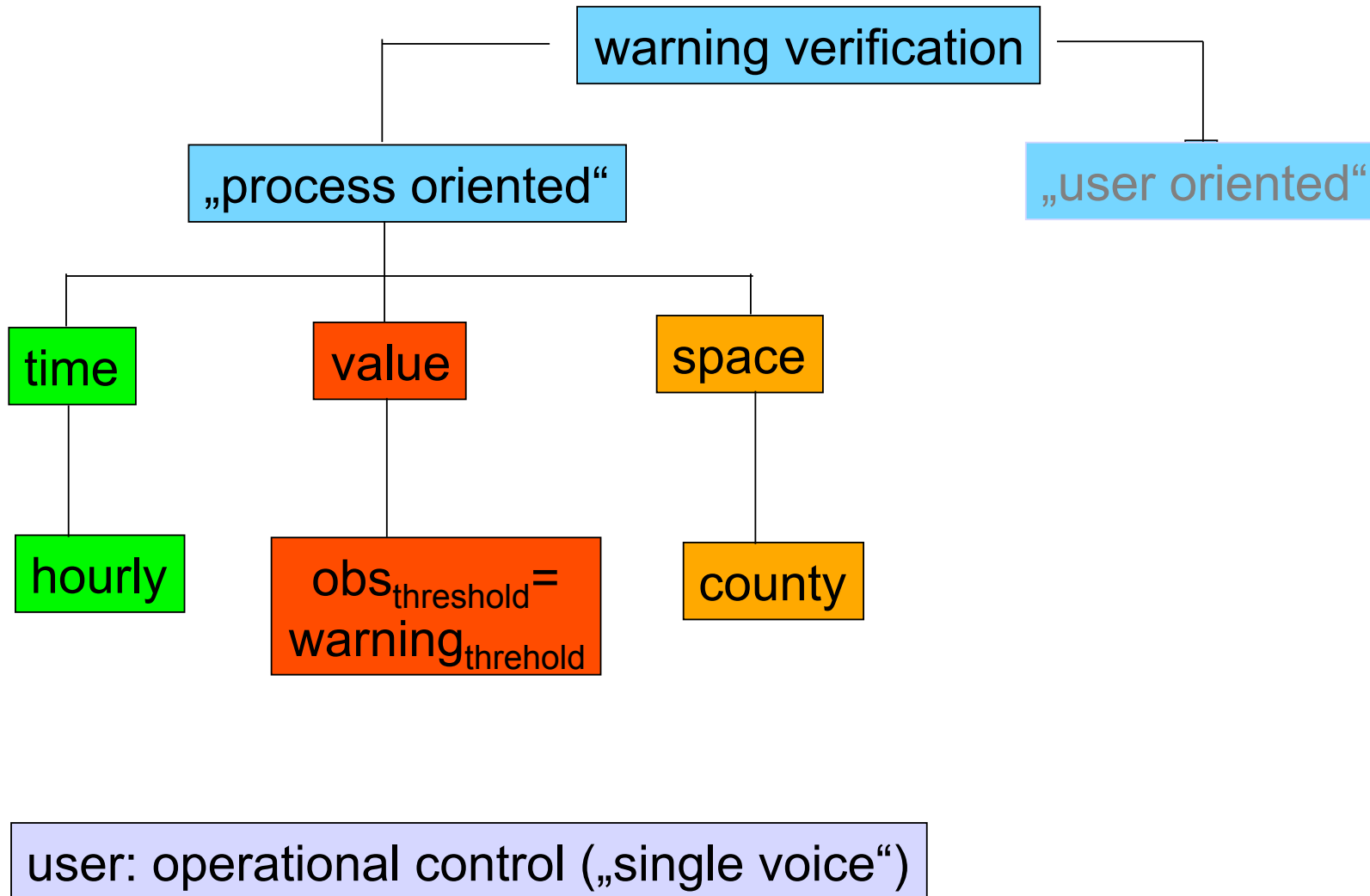
- hourly (SYNOPS), e.g. NCEP, UKMO, DWD as “process oriented verification”
- “events”:
 - warning and/or obs immediately followed by warning
 - obs in an interval starting at first threshold exceedance (e.g. UKMO 6 hours before the next event starts)
 - even “softer” definition: as “extreme events”
-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



Issue: matching warning and obs

temporal

6. How do you match observations and warnings ? What is the actual lead time (as opposed to officially desired) which you require to count a warning as a "hit"?			
	yes	no	Response Count
on hourly basis	20.0% (1)	80.0% (4)	5
on three hourly basis	50.0% (3)	66.7% (4)	6
as "events"	81.8% (9)	18.2% (2)	11
lead time of at least 1 hour	50.0% (2)	50.0% (2)	4
lead time of least 2 hours	40.0% (2)	60.0% (3)	5
lead time of 2 or more hours	75.0% (6)	25.0% (2)	8
	<i>answered question</i>		16
	<i>skipped question</i>		3





Warning verification

„process oriented“

(er) e en orien ed

time/
events

value

space

- 1. warning
- 2. obs intervals

hit

false alarm

radius

region

$\text{delta}_{\text{intensity}} = 0$

$\text{delta}_{\text{intensity}} > 0$

user: emergency services

user: media



Issue: matching warning and obs

hit

								hourly, "process oriented" verification	"event oriented" verification
time	15	16	17	18	19	20	21		
observation			1					1 hit	1 hit
warning		1	1	1	1			3 false alarms	
time of issue		X							

**miss (too late)
or
hit (still useful)**

								hourly, "process oriented" verification	"event oriented" verification
time	15	16	17	18	19	20	21		
observation		1						1 miss (or hit)	1 miss
warning		1	1	1				2 false alarms	
time of issue		X							

**hit
+
false alarm
(too long)**

								hourly, "process oriented" verification	"event oriented" verification
time	15	16	17	18	19	20	21		
observation			1					1 hit	1 hit
warning		1	1	1	1	1		2 false alarms	(including 1 false alarm)
time of issue		X							



Issue: matching warning and obs

Large difference of model verification!

spatial

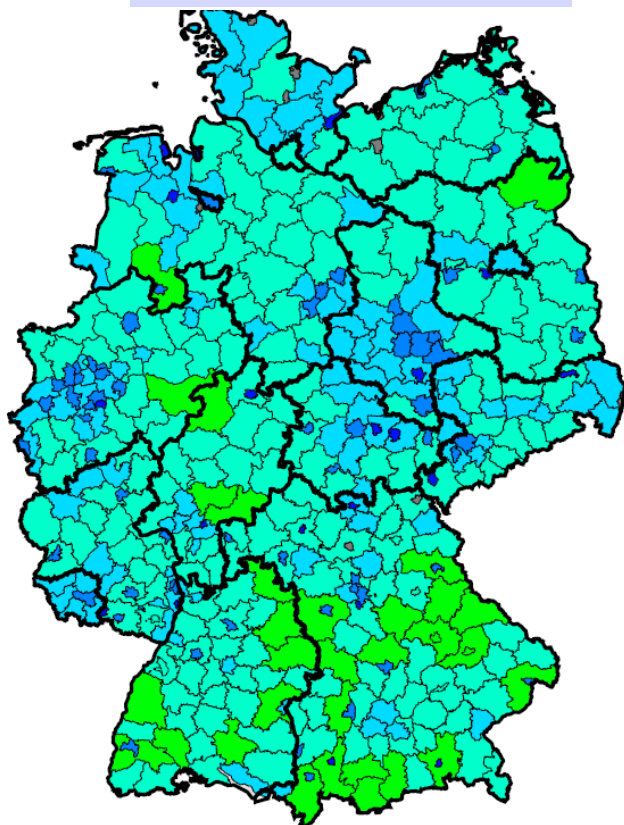
- sometimes “by-hand” (e.g. Switzerland, France)
- worst thing in the area
- dependency on area size possible
- “MODE-type” (**M**ethod for **O**bject-based **D**iagnostics **E**valuation)



Issue: matching warning and obs

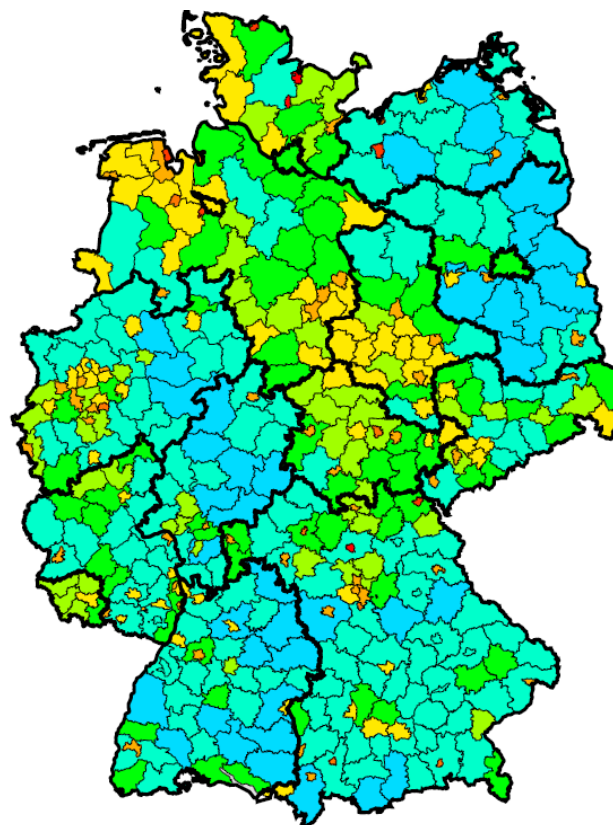
Thunderstorms

Base rate / h



- ≤ 0.001
- ≤ 0.002
- ≤ 0.005
- ≤ 0.010
- ≤ 0.020
- ≤ 0.050
- ≤ 0.100
- ≤ 0.150
- ≤ 0.200
- ≤ 0.300
- ≤ 0.500
- > 0.500
- no Obs

bias



- ≤ 0.500
- ≤ 1.000
- ≤ 2.000
- ≤ 4.000
- ≤ 6.000
- ≤ 8.000
- ≤ 10.000
- ≤ 20.000
- ≤ 50.000
- ≤ 80.000
- ≤ 150.000
- > 150.000
- no Obs



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



Issue: measures

7. Which measures do you use to summarise the quality ? Check all that apply.

	Response Percent	Response Count
hit rate (= probability of detection)	83.3%	10
false alarm rate (percentage of falsely warned non-events)	58.3%	7
false alarm ratio (percentage of false warnings)	25.0%	3
TS (threat score)	50.0%	6
ETS (equitable threat score)	8.3%	1
HSS (Heidke skill score)	41.7%	5
HKS (Hansen-Kuippers score = TSS True skill score)	16.7%	2
ROC	16.7%	2
metric taking costs and losses explicitly into account	8.3%	1
Hide replies Other (please specify)		4
<p>1. Dont now Fri, Jun 5, 2009 9:42 AM Find...</p> <p>2. I dont know if they compute any of these measures Thu, Jun 4, 2009 9:57 PM Find...</p> <p>3. W don't use any of deal above, but we use another criteria (for example, Obukhov criteria). Thu, Jun 4, 2009 4:39 PM Find...</p> <p>4. odds ratio Thu, Jun 4, 2009 12:24 PM Find...</p>		
	<i>answered question</i>	12
	<i>skipped question</i>	7



Issue: measures

Finley dataset, 1884

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

- “everything” used (including Extreme Dependency Score EDS, 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 frequently used, since definition of the no-forecast/no-obs category problematic
- 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
- “F-measure”

$$F_{\beta} = (1 + \beta^2) * \frac{POD * (1 - FAR)}{\beta^2 * POD + 1 - FAR}$$

“After years of study we ended up in using the value 1.2 for β for extreme weather....”



Issue: “Interpretation” of results

8. Do you have performance targets for warnings and are there consequences because of the results?		Response Percent	Response Count
yes		23.1%	3
no		76.9%	10
Hide replies Other (please specify)			1
1. i am not sure Fri, Jun 5, 2009 9:43 AM Find...			
		<i>answered question</i>	13
		<i>skipped question</i>	6



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. 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*
- 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 C/L ratio into account



Issue: “Interpretation” of results

Consequences

- changing forecasting process
 - e.g shortening of warnings at DWD dramatically reduced false alarm ratio based on hourly verification almost without reduction in POD
 - creating new products (probabilistic forecasts)



Issue: user-based assessments

9. Do you do user based assessments? What questions do you ask?		Response Percent	Response Count
yes		50.0%	7
no		50.0%	7
		Hide replies Other (please specify)	3
<ol style="list-style-type: none"> i am not sure Snow depth critical for traffic, hydrological risk Did you have sufficient lead time? Did you take any precautions? Was the infomation useful? 			
			14
			5



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, focus groups, user workshops, public opinion monitoring, feedback mechanisms, anecdotal information
- presentation of warnings to the users essential
- “vigilance evaluation committee” (Meteo France /Civil Authorities)
- 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?
 -



Comparing warning guidances and warnings

Example here, gust warnings

- Warning guidance: "Local model gust forecast" (=mesoscale model)
- warning: human (forecaster)

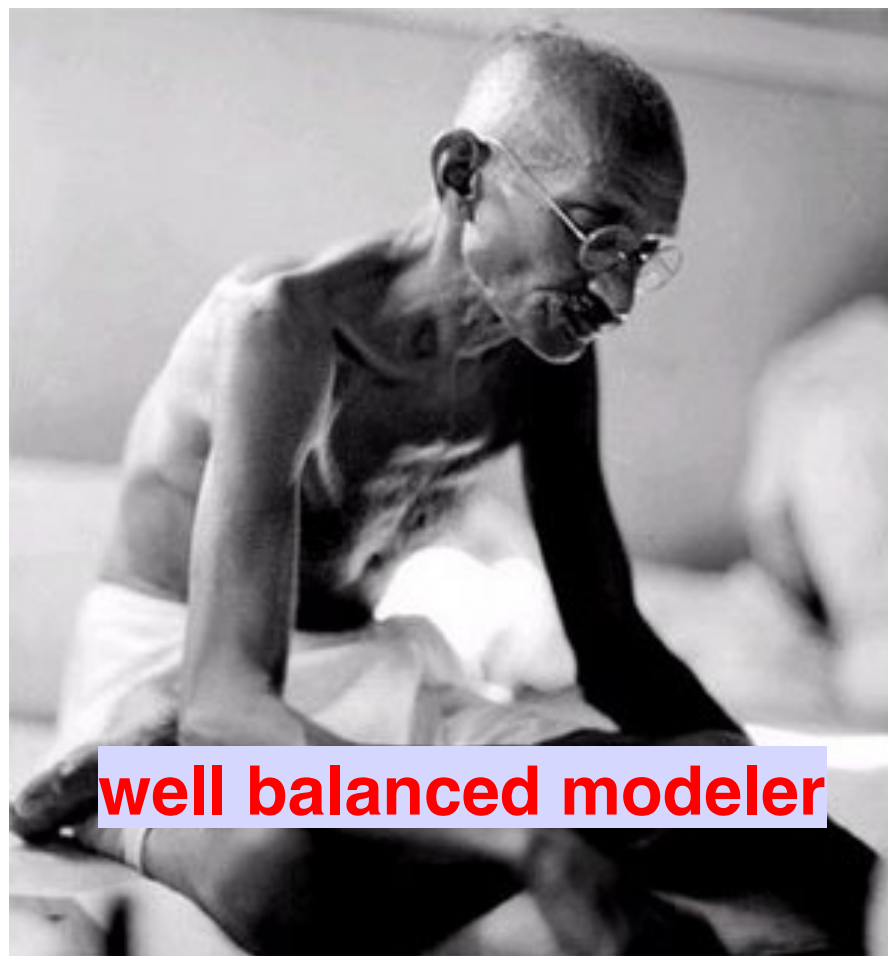
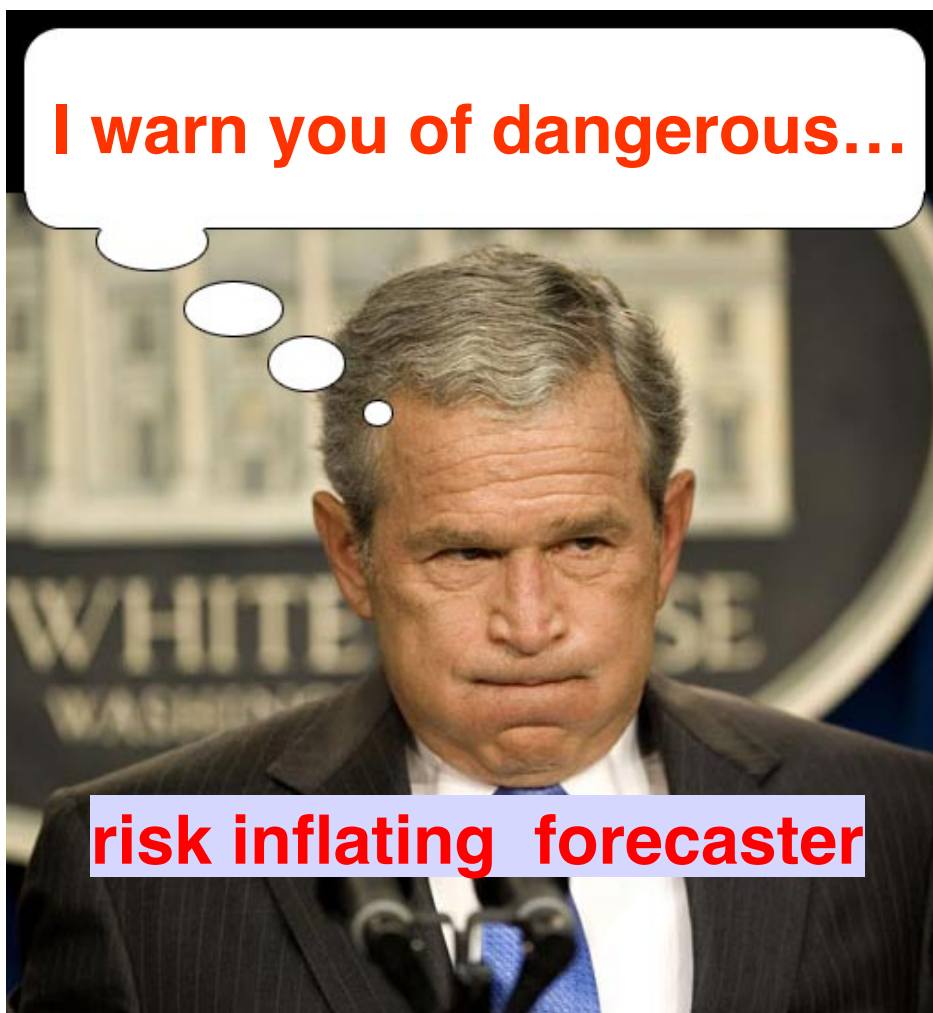


Comparing warning guidances and warnings

10. Do you compare warning guidance systems with warnings?			Response Percent	Response Count
no			46.7%	7
model guidances			53.3%	8
statistical products			20.0%	3
expert systems			13.3%	2
			<i>answered question</i>	15
			<i>skipped question</i>	4



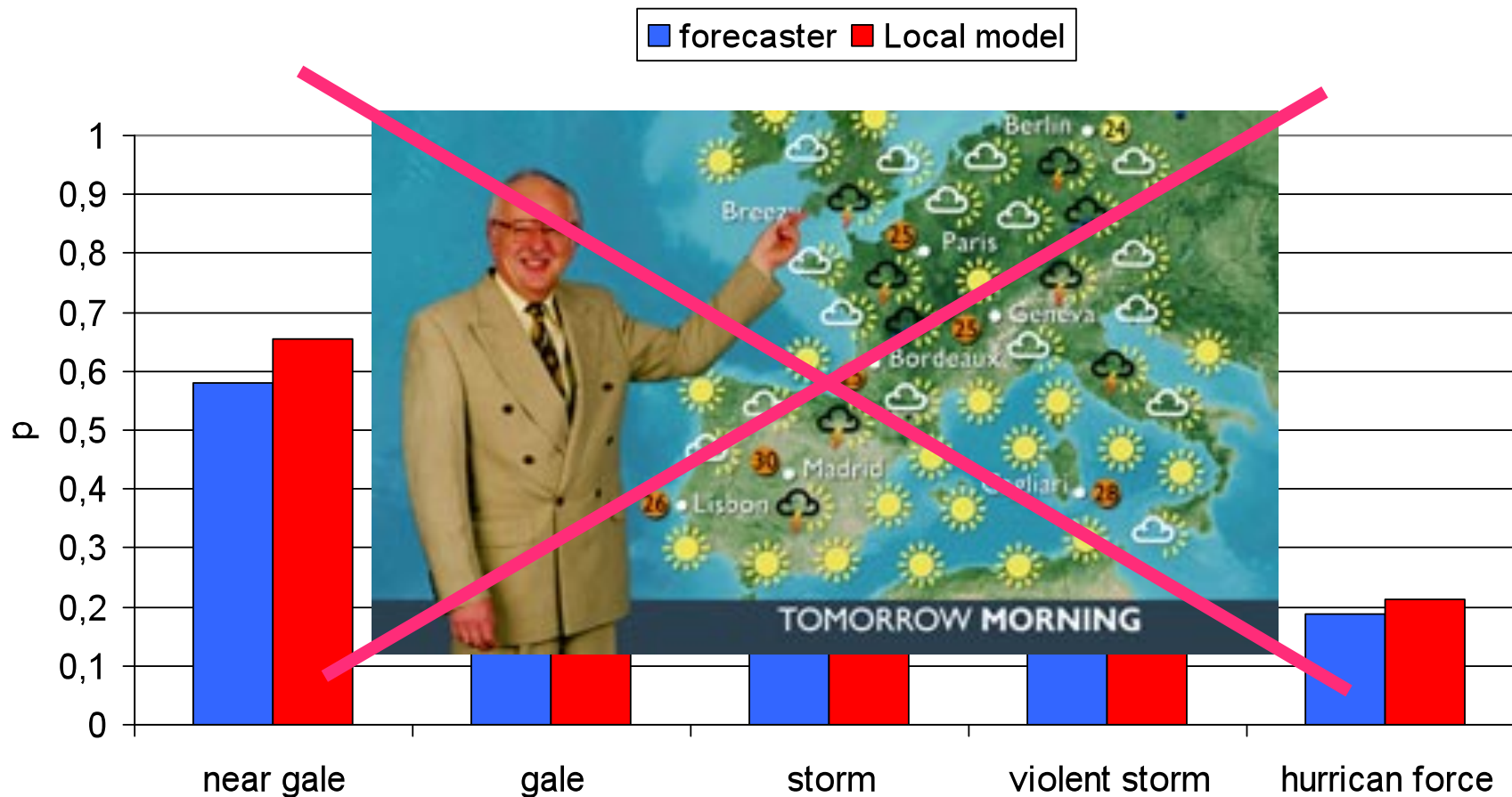
Comparing warning guidances and warnings





Issue: Comparing warning guidances and warnings

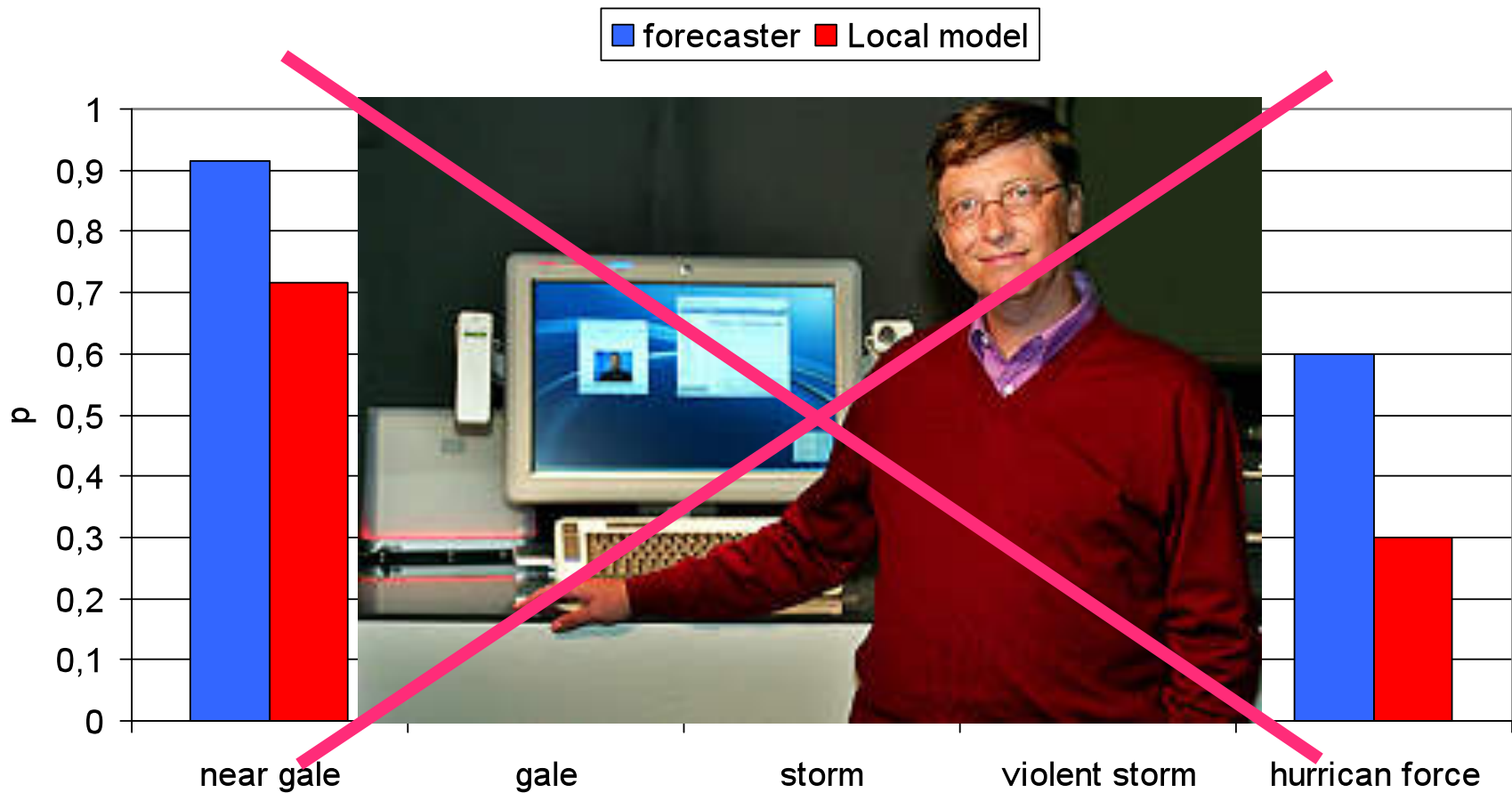
Heidke skill score





Issue: Comparing warning guidances and warnings

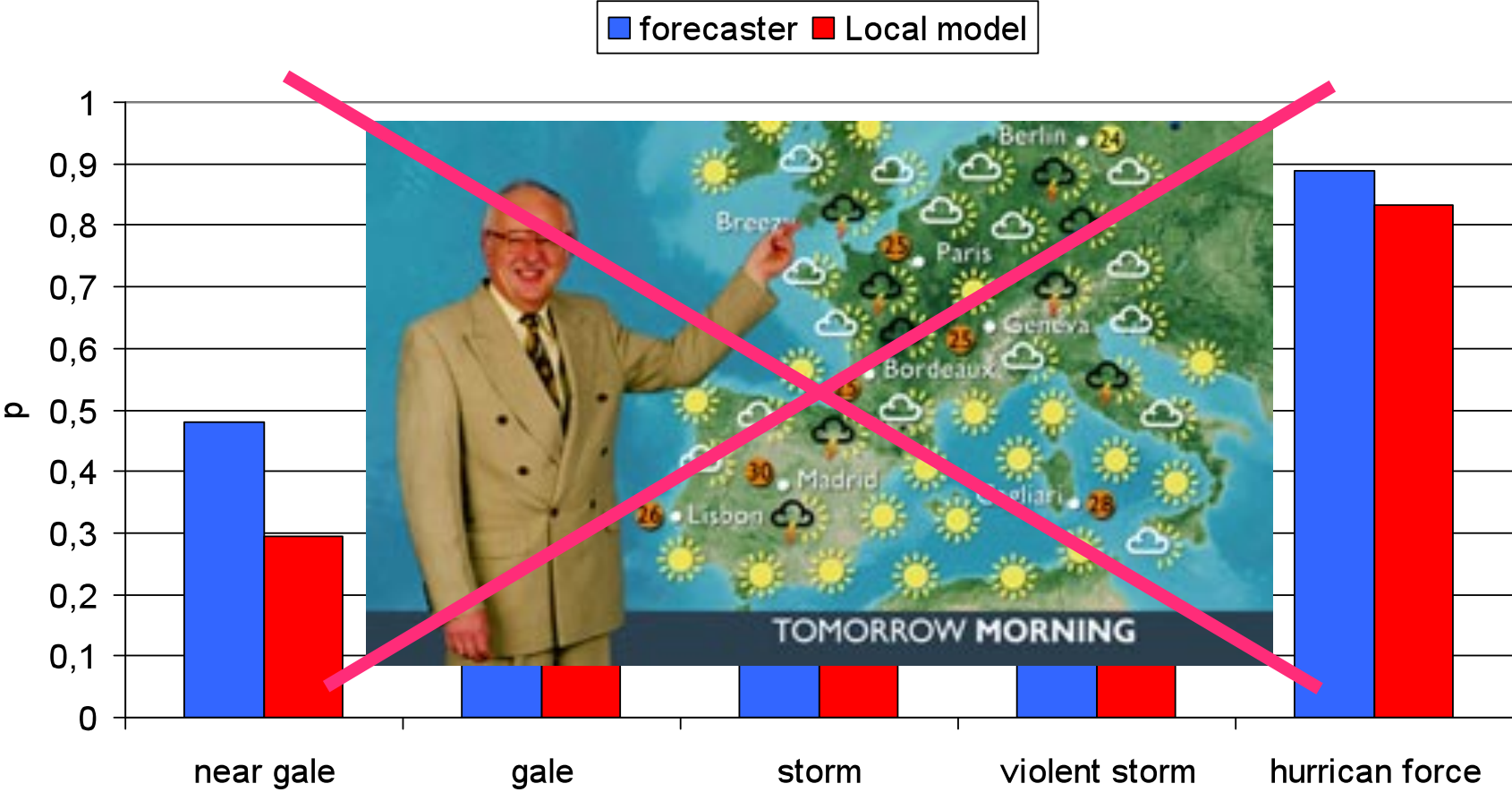
hit rate





Issue: Comparing warning guidances and warnings

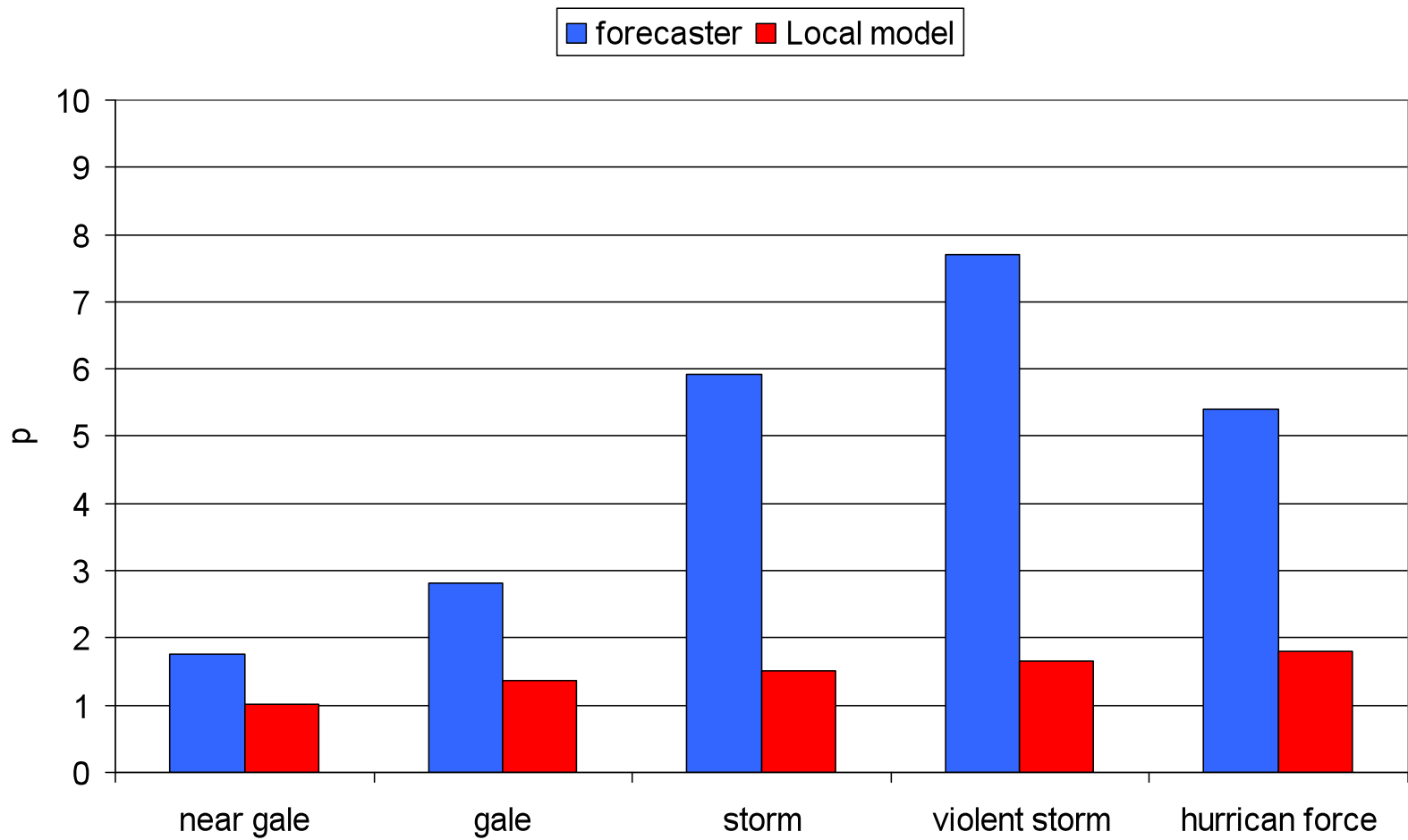
false alarm ratio





Issue: Comparing warning guidances and warnings

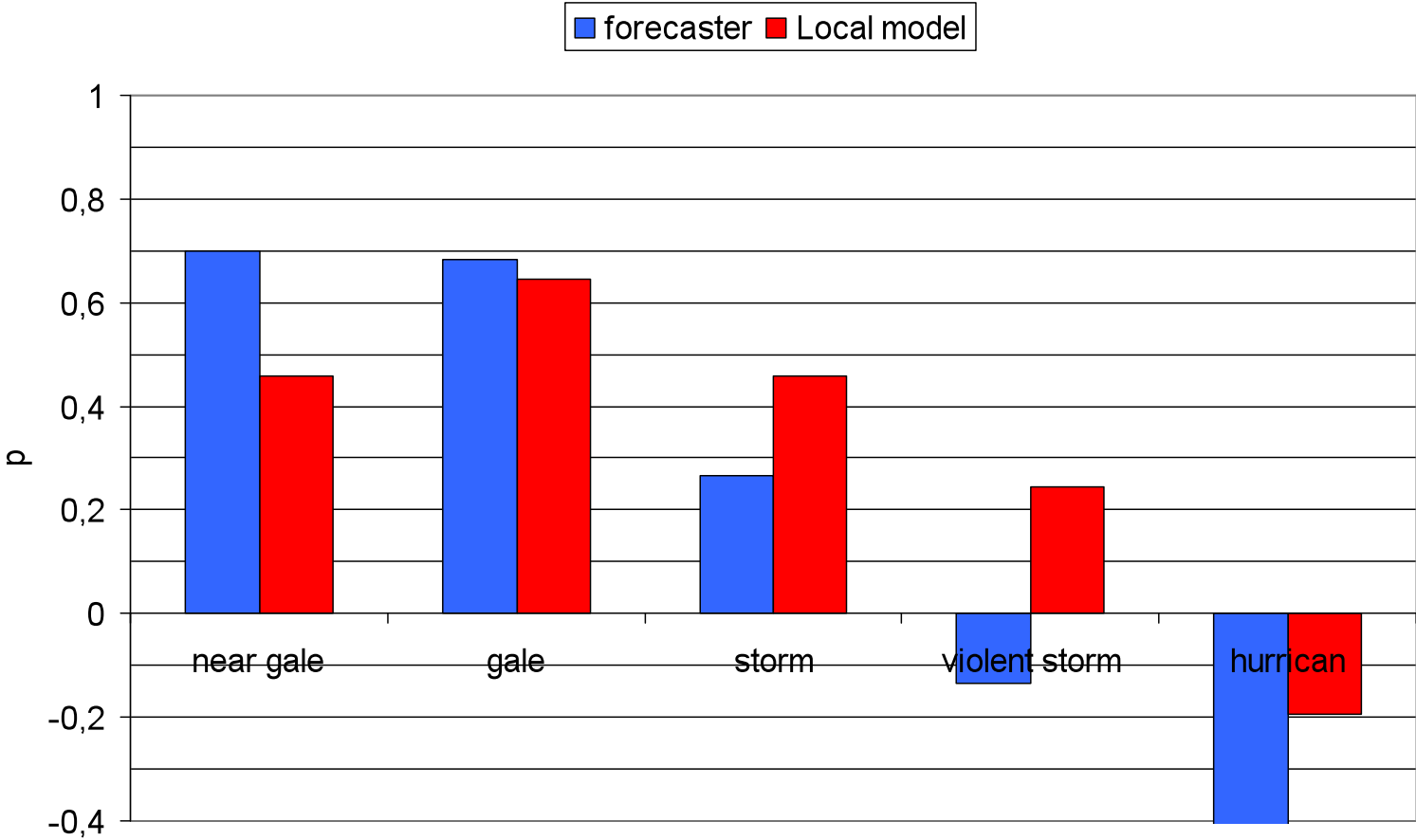
Bias





Issue: Comparing warning guidances and warnings

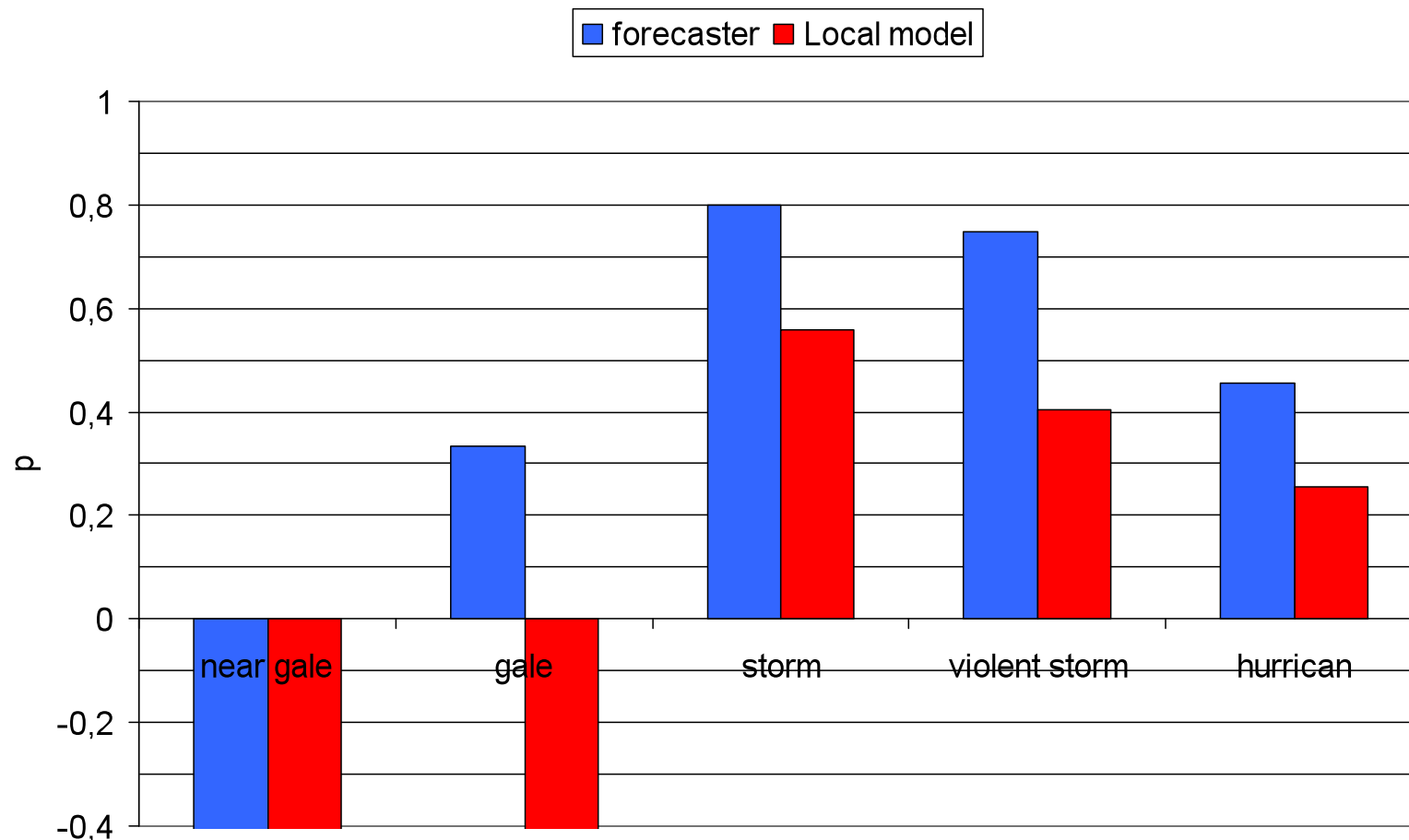
relative value for C/L=0,1





Issue: Comparing warning guidances and warnings

relative value for $C/L=0,01$





Issue: Comparing warning guidances and warnings

very different biases
→ comparison of apples and oranges

But is there a way of "normalising",
so that at least the **po en ial** can be compared ?



Issue: Comparing warning guidances and warnings

Re-calibration
„model bias = forecaster bias“
 $cdf(model) = cdf(forecaster)$

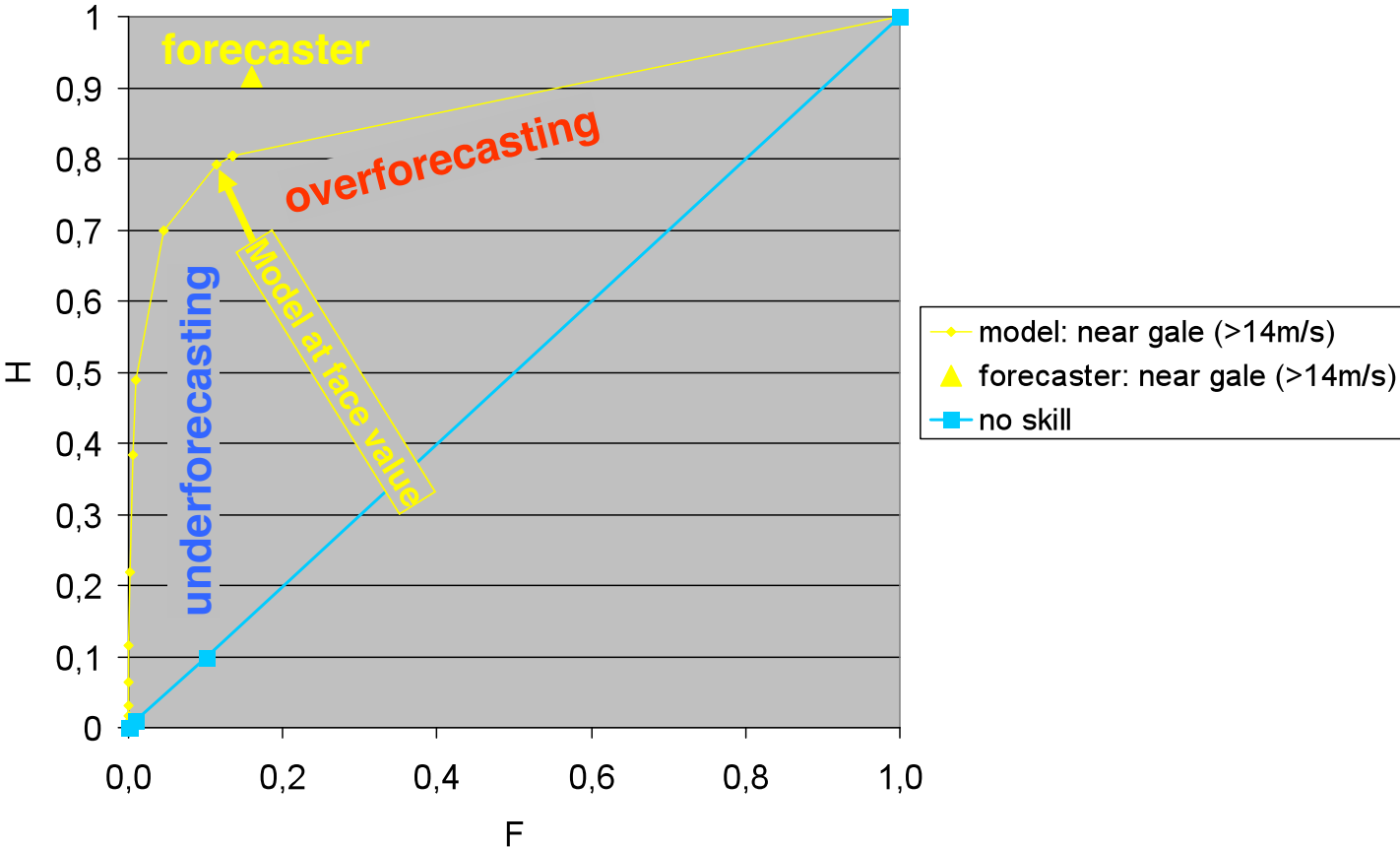
model in m/s	---->	„model gust interpretation for warnings “
13	---->	14 (near gale)
16	---->	18 (gale)
22	---->	25 (storm)
25	---->	29 (violent storm)
30	---->	33 (hurricane force)

Verification of heavily biased model ? **Quite similar to forecaster !**



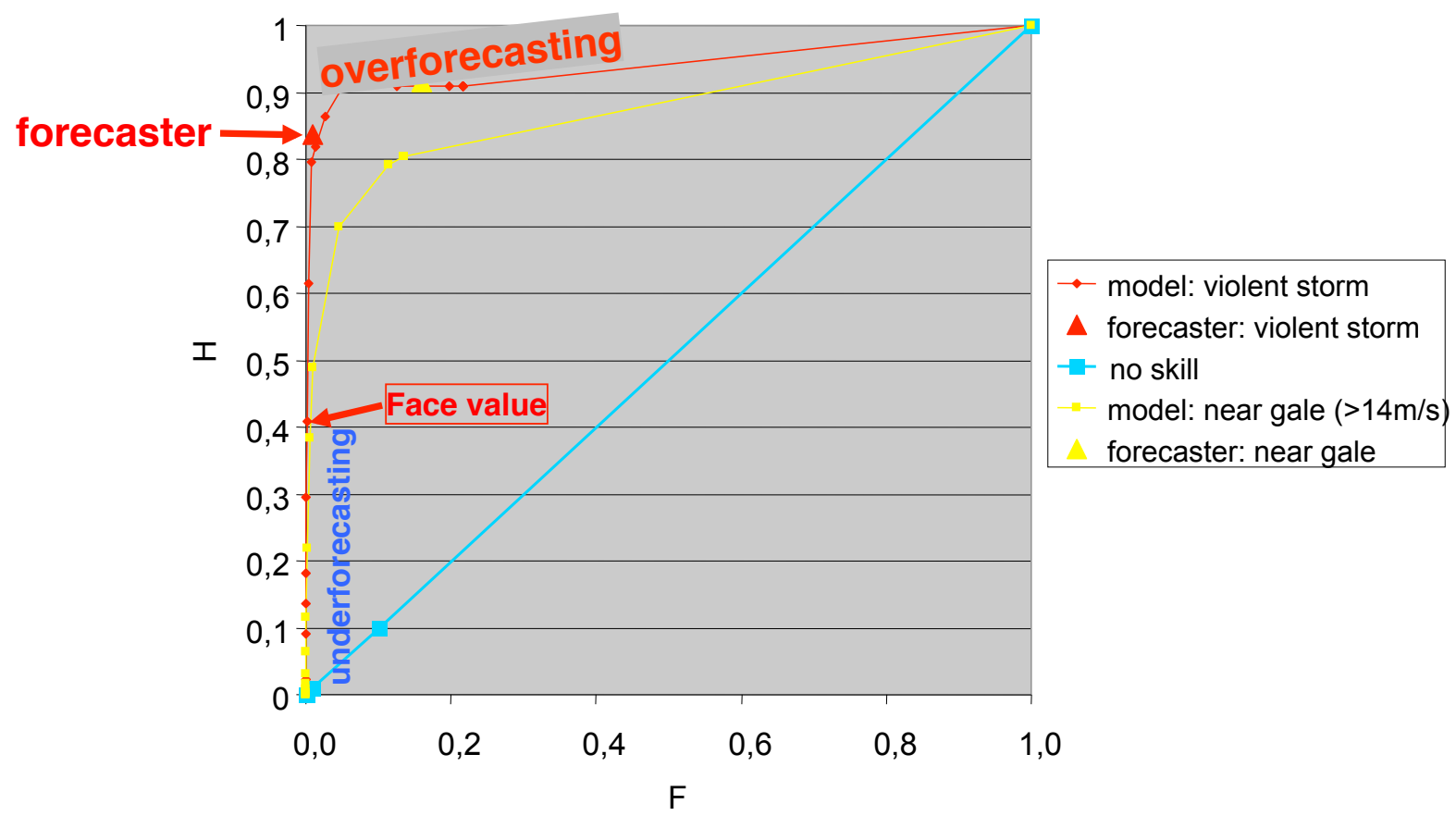
Issue: Comparing warning guidances and warnings

Relative Operating Characteristics (ROC)





Issue: Comparing warning guidances and warnings





Issue: Comparing warning guidances and warnings

Conclusions for comparative verification man vs machine

End user verification: verify at face value

Model (guidance) verification: measure **potential**



Summary

Users of warning are either direct and handling verification is also either direct.

Each choice of a parameter of the verification method has to be well oriented here in no one is efficient.



Can we earn a better?



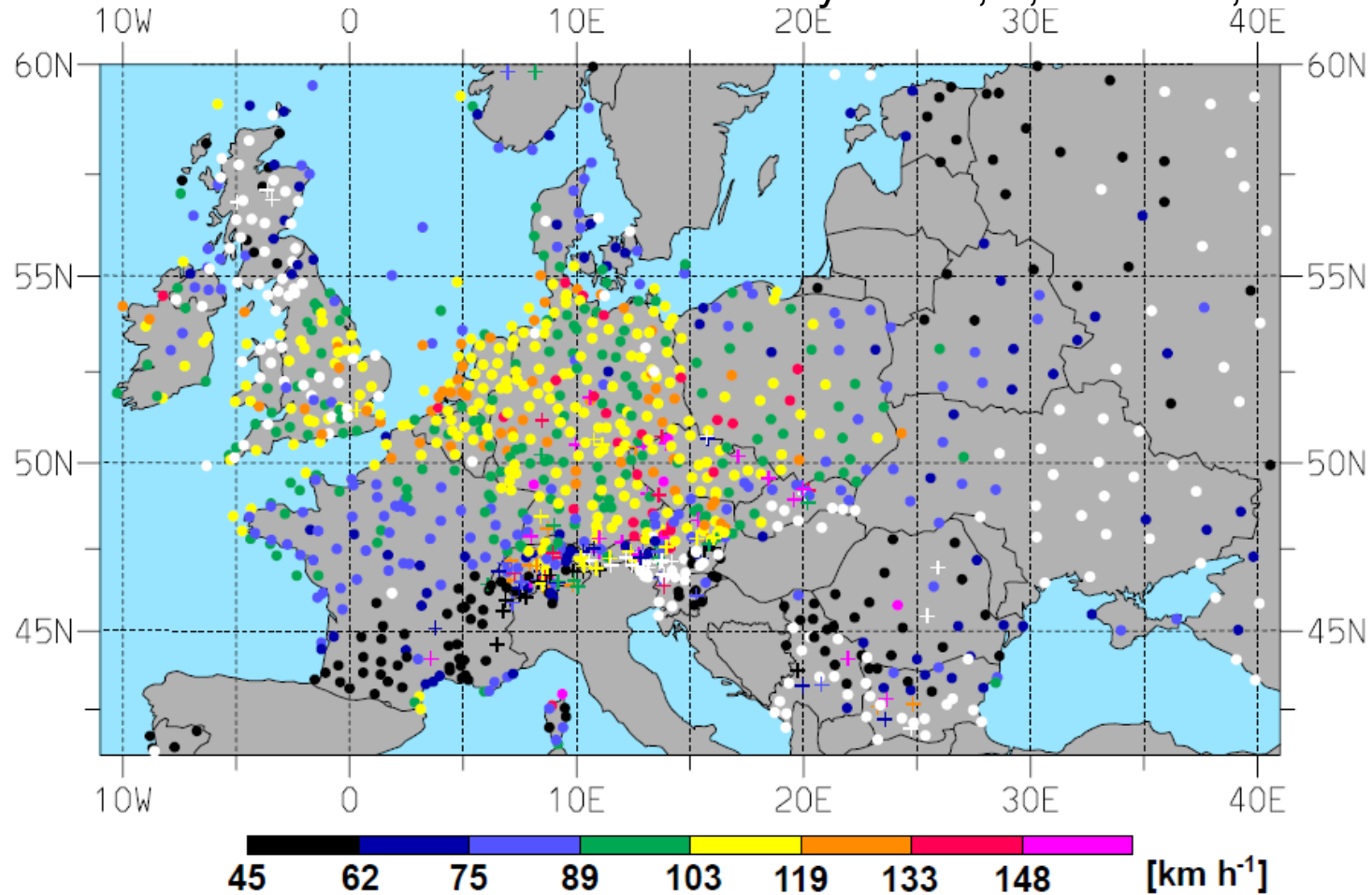


Fig. 6. Maximum wind gusts (in km h^{-1}) at different synoptic stations reported during the period from 00:00 UTC 17 January to 18:00 UTC 19 January 2007. Dots (crosses) delineate lowland (mountain) stations. Lowland stations possess an altitude lower than 800 m a.s.l. White symbols denote stations where no wind gusts were observed or reported.