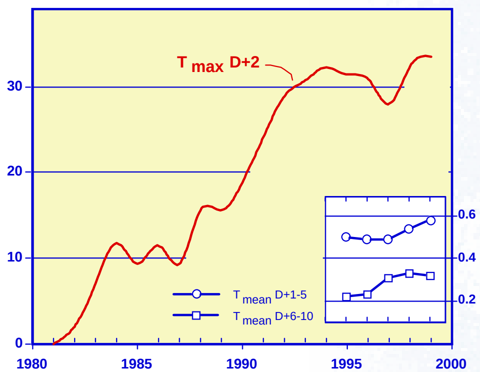
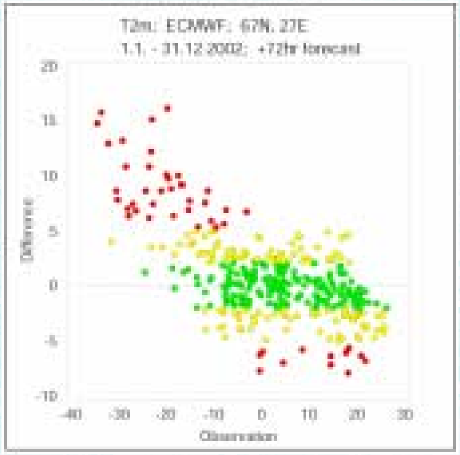
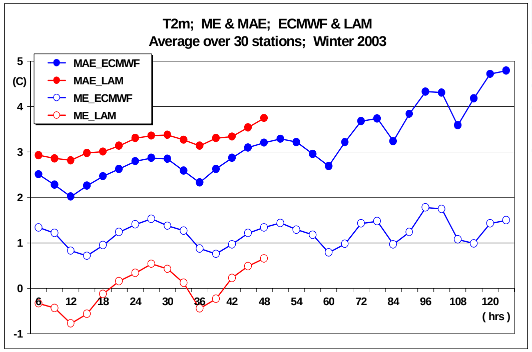


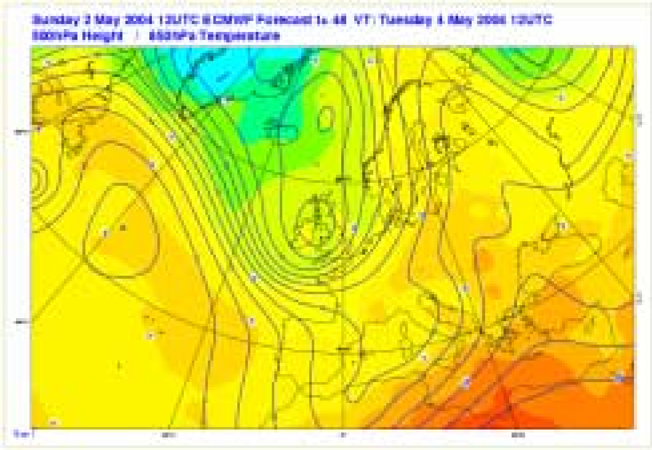


Pertti Nurmi

(Finnish Meteorological Institute)

Design of Operational Verification Systems





Montreal
17.9.2004

pertti.nurmi@fmi.fi

Acknowledgements:

- ✓ **HUNGARY**: (!) *Gabriella Csima, Istvan Ihasz, Helga Toth, Ervin Zsótér*
- ✓ **AUSTRALIA**: *Beth Ebert et al.*
- ✓ **FRANCE**: *Frederic Atger, Bernard Strauss*
- ✓ **GERMANY**: *Ulrich Damrath, Martin Göber, Thomas Kratzsch*
- ✓ **NORWAY**: *Helen Korsmo*
- ✓ **SLOVENIA**: *Miha Razinger*
- ✓ **U.S.A.**: *Barb Brown, Keith Brill, Ying Lin, Andy Loughe*

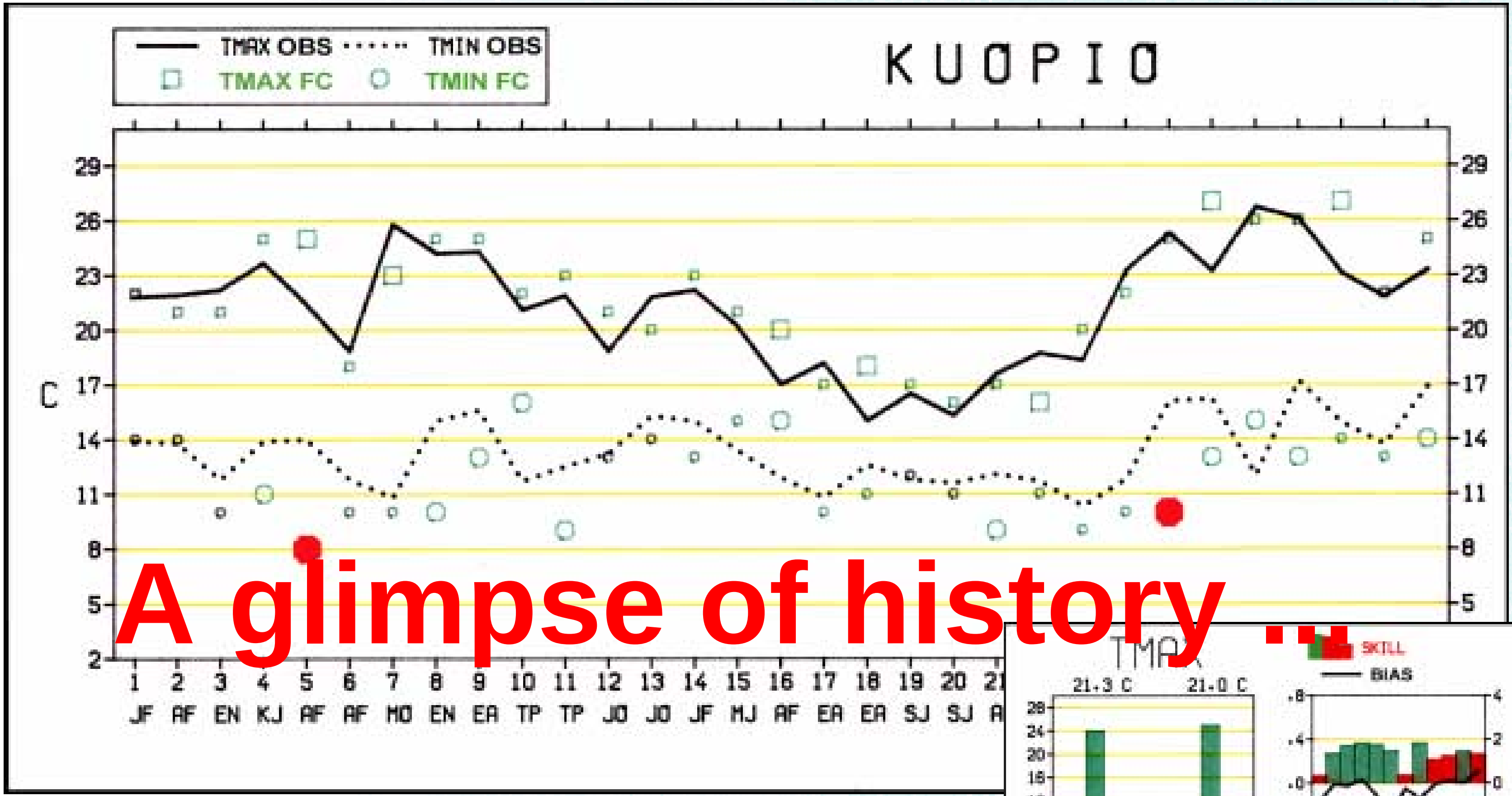
\+ **ECMWF**

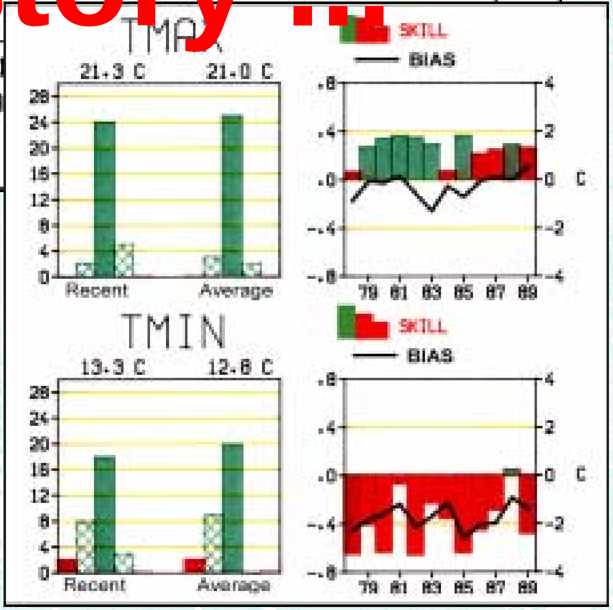
HUNGARIAN METEOROLOGICAL SERVICE

Forecast Verification and Development Division

OMSZ

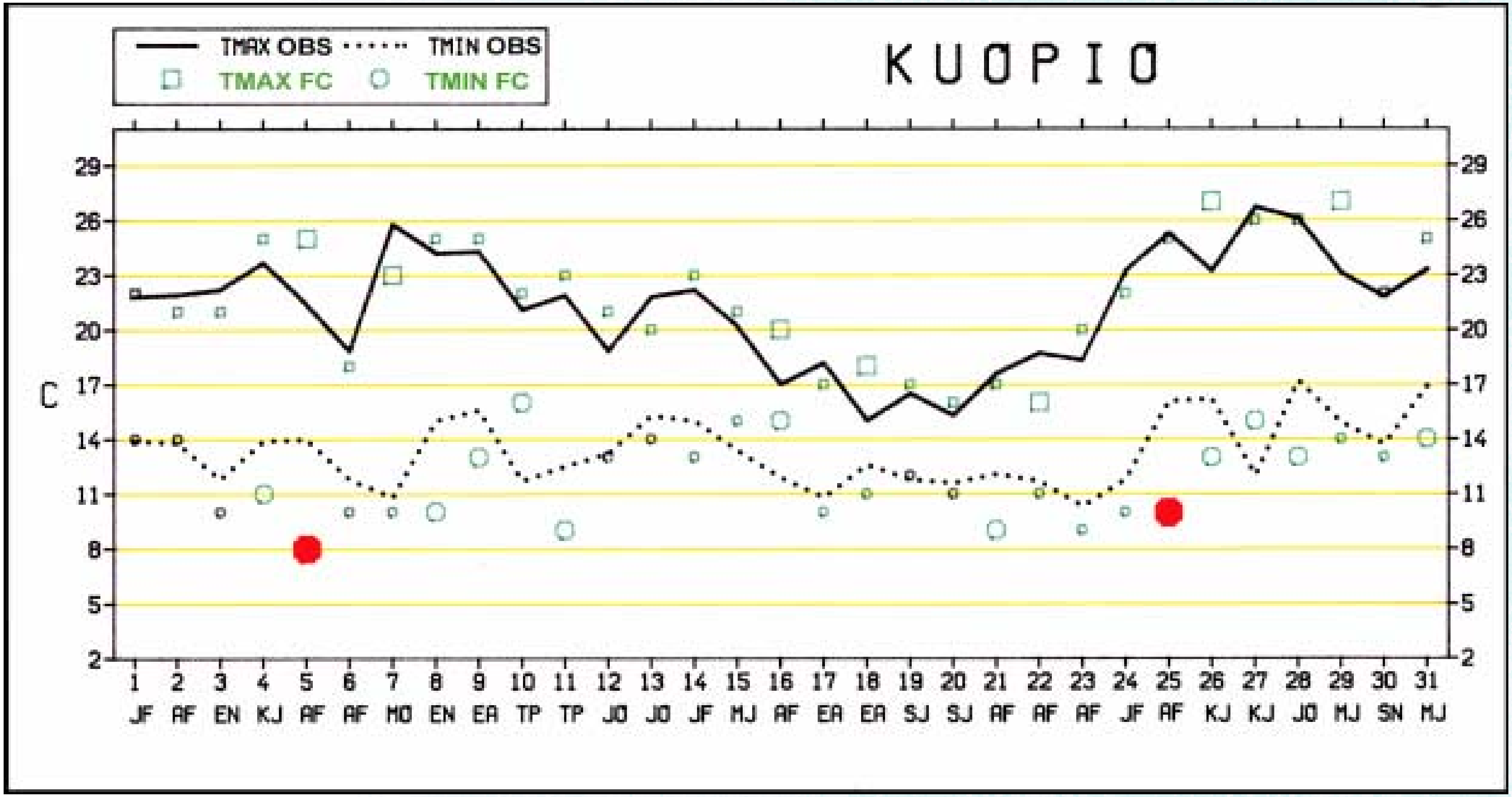
A glimpse of history ...



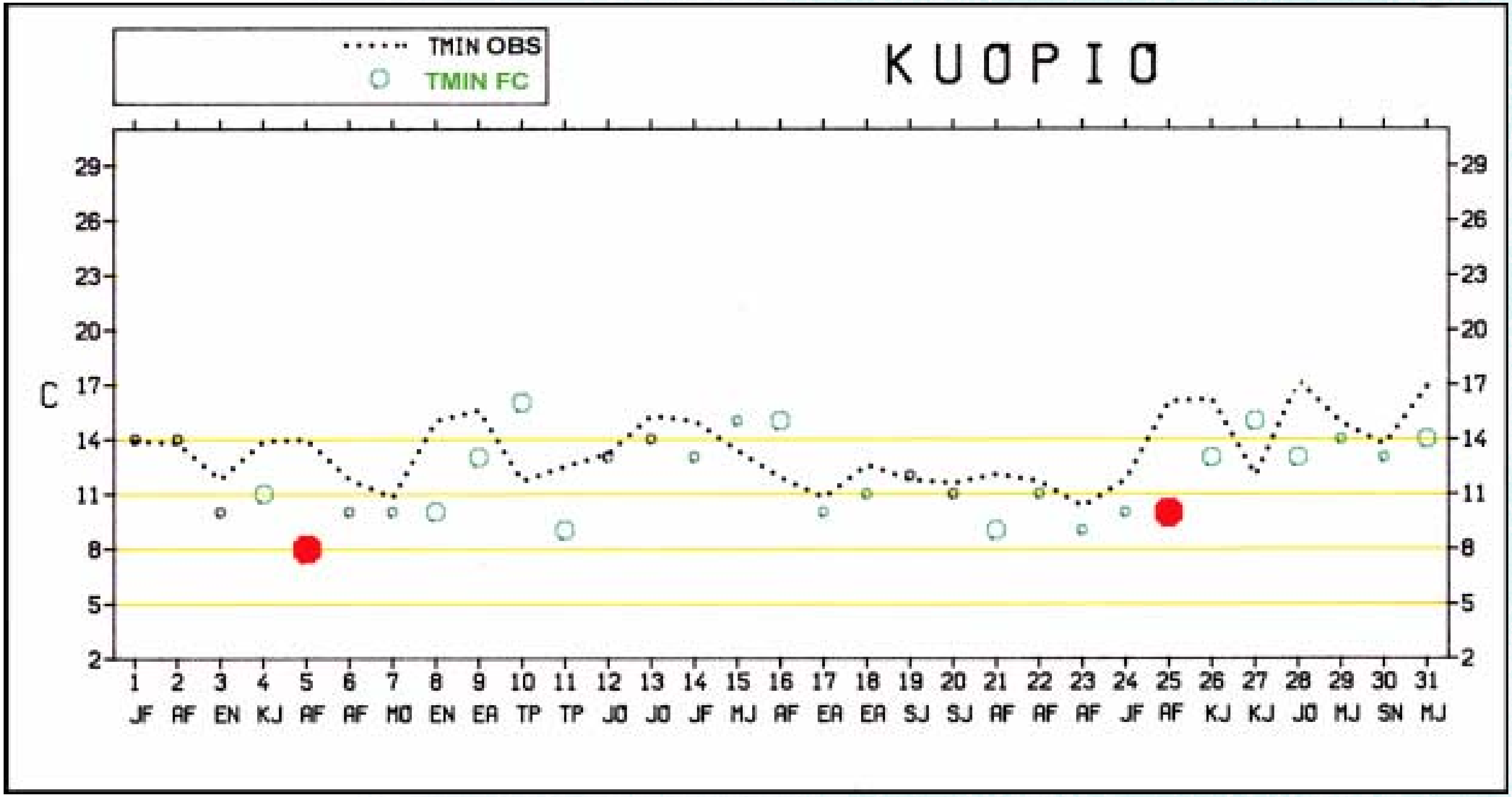


From: FMI operational verification system, of late 1980s to early 1990s, i.e. before PC graphics or Internet were "invented"

Monthly printed statistics →

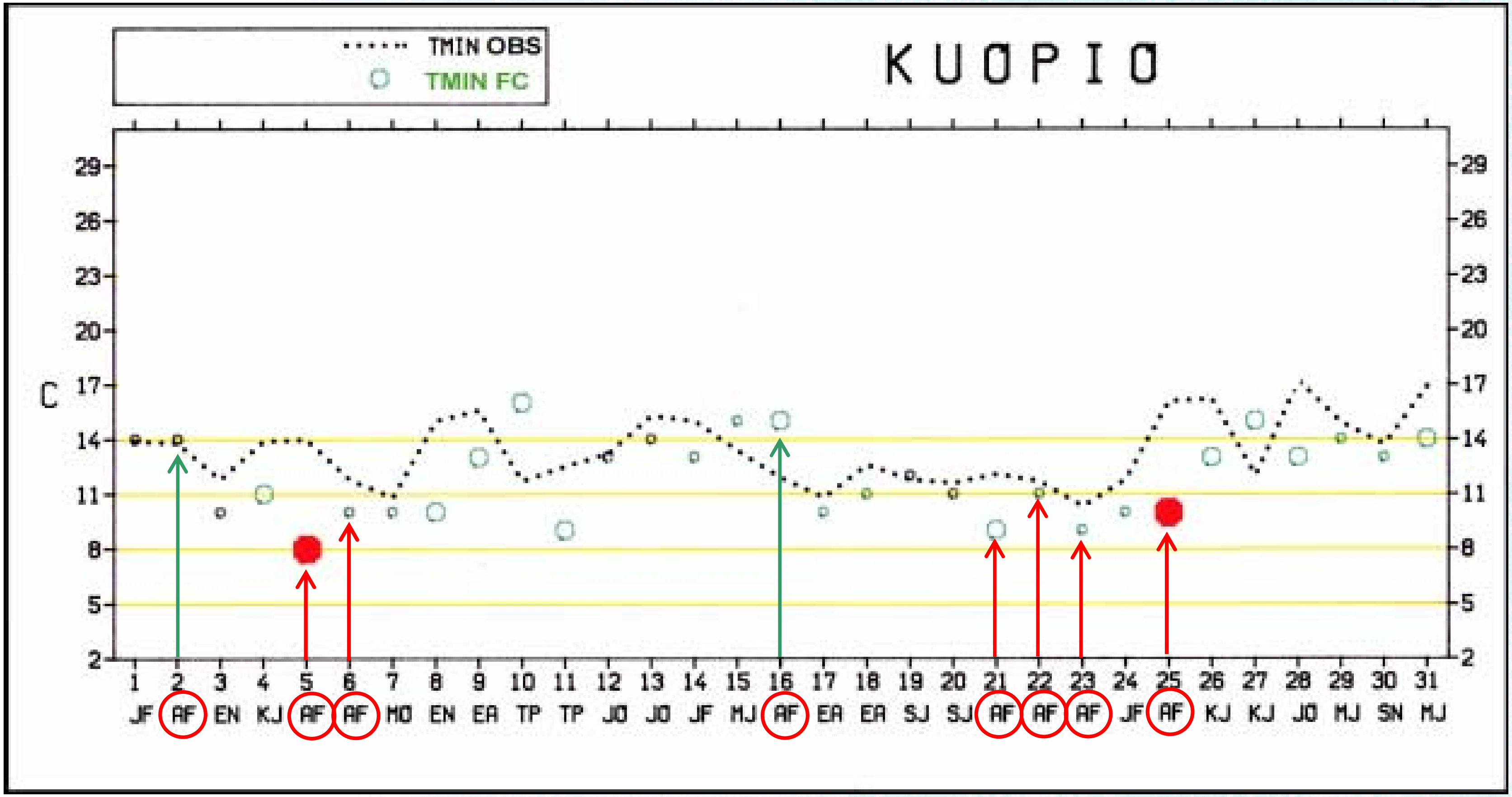


From: FMI operational verification system, July 1989



From: FMI operational verification system, July 1989

Cold bias: – 1.5 ^{o}C

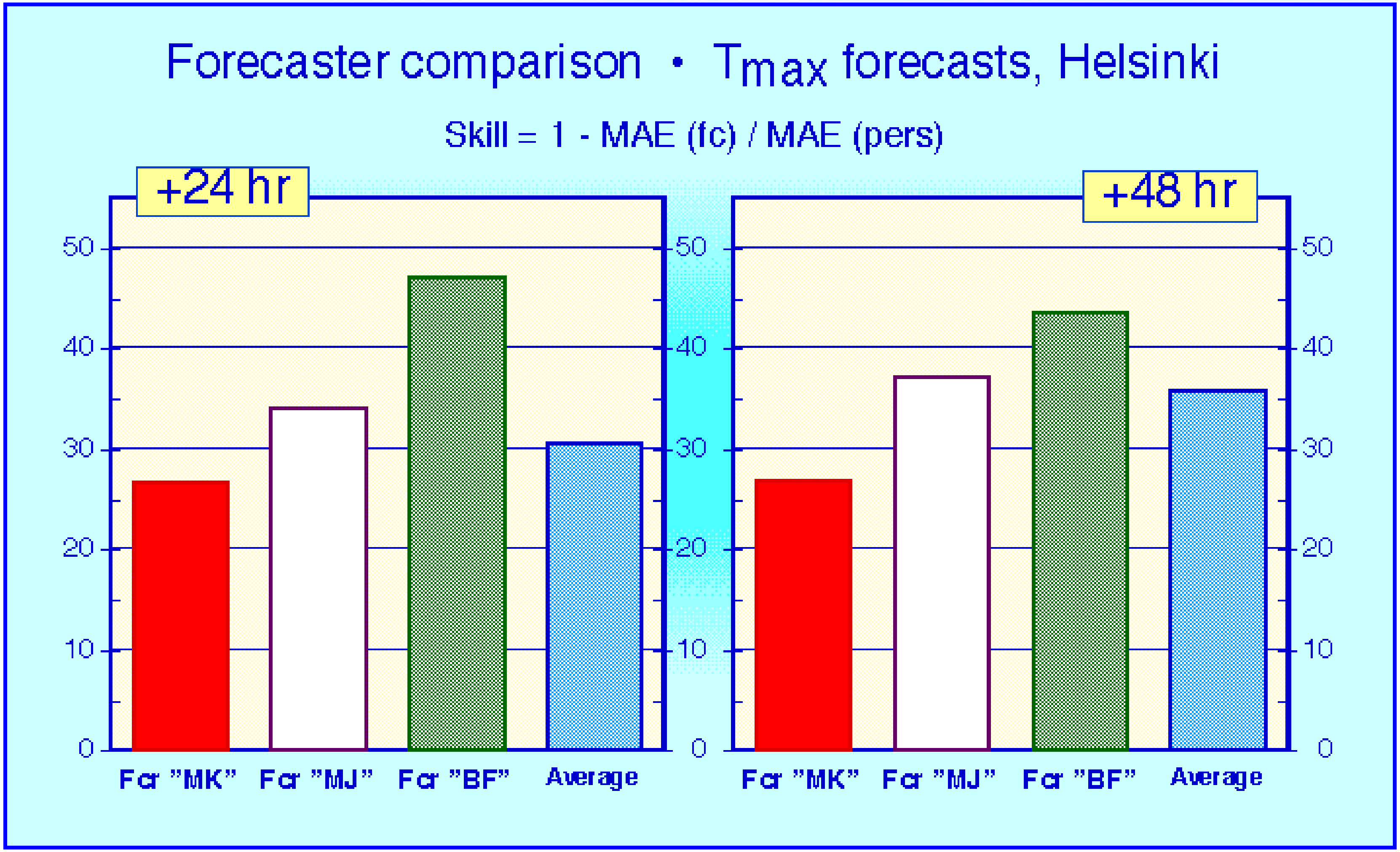


From: FMI operational verification system, **July 1989**

Cold bias: – 1.5 ^{o}C

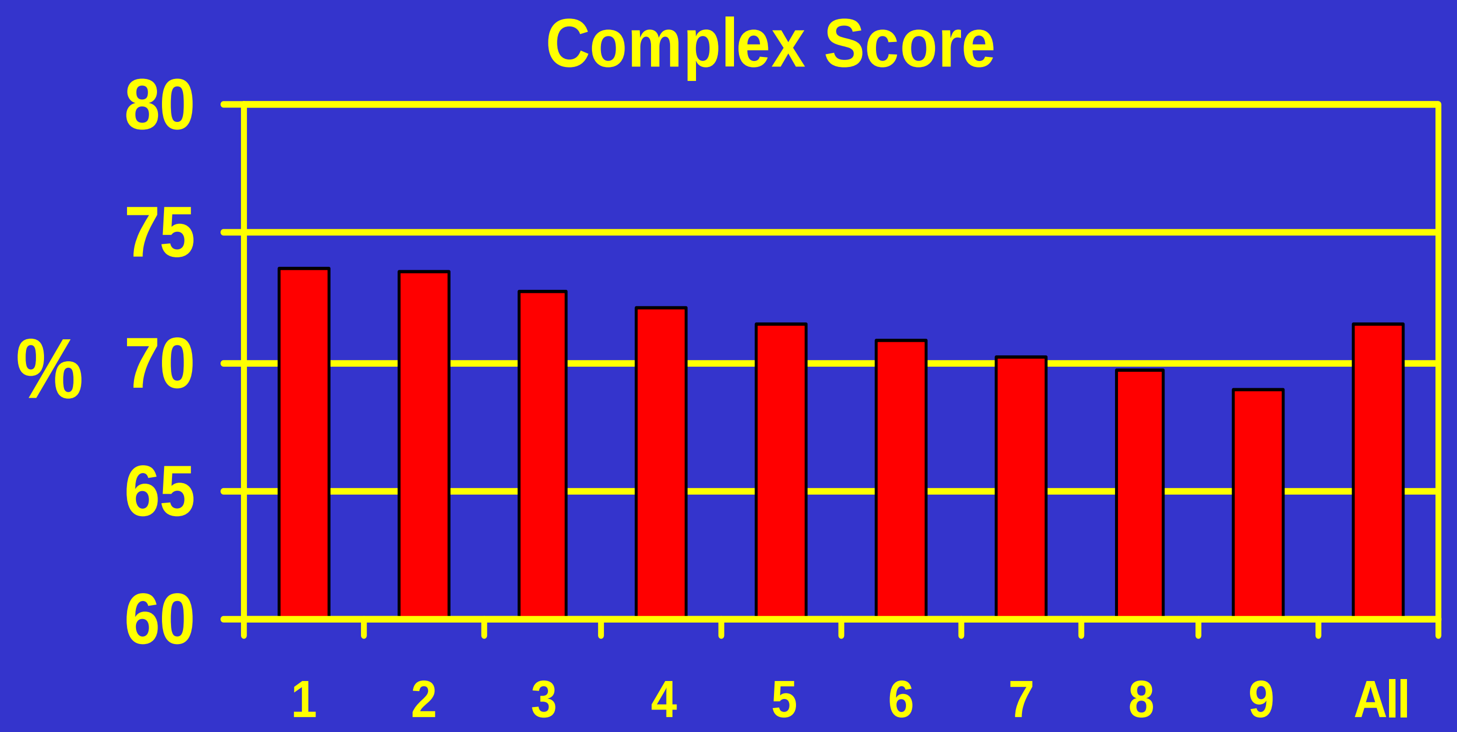
Forecaster AF´s cold bias: – 2.5 ^{o}C

Individual verification, Summary Statistics



Case, HUNGARY *(Zsótér)*

Verification per forecasters, Jan – Dec 2000, Budapest



- ”Gently stimulate to improve forecast quality”

Case, NORWAY *(Korsmo)*

Personal verification feedback by e-mail !

Case, NORWAY *(Korsmo)*

Personal verification feedback by e-mail !

File Edit View Go Message Tools Window Help

Subject: Verifikasjon for 2004-05-23 12:00:00 **From:** helen.korsmo@met.no 05/23/2004 03:00 PM

Avistabell for 23/05-2004, kl. 12:00, skrevet av magnuso, parameter: T2m

STED	OBS	SUBJ:12+24	SUBJ.SCORE	HIRLAM20: 00+36	HIRLAM10: 00+36	ECMWF 12+48
Alta	7	7	BLINK!	5	7	7
Andøya	6	6	BLINK!	6	6	7
Bardufoss	6	7	BLINK!	8	8	9
Bodø	7	7	BLINK!	7	7	8
Brønnøysund	8	8	BLINK!	7	8	8
Glomfjord	6	8	Bra!	8	8	8
Hammerfest	5	4	BLINK!	4	5	4
Karasjok	6	5	BLINK!	4	4	5
Kautokeino	6	6	BLINK!	4	6	7
Kirkenes	3	3	BLINK!	5	4	2
Lakselv	7	6	BLINK!	4	5	6
Longyearbyen	2	2	BLINK!	0	1	2
Mosjøen	7	9	Bra!	9	8	9
Nord-Solver	9	8	BLINK!	7	6	8
Nordreisa	6	7	BLINK!	7	7	10
Rustefjelbma	4	4	BLINK!	5	4	4
Røst	7	7	BLINK!	6	6	7
Saltdal	7	10	Oups..	10	11	10
Sletnes fyr	3	4	BLINK!	5	5	4
Sortland	5	7	Bra!	7	7	7
Svolver	7	7	BLINK!	7	7	7
Torsvåg fyr	6	5	BLINK!	6	6	6
Tromsø	6	6	BLINK!	7	7	7
Vadsø	3	4	BLINK!	5	5	2
Vardø	4	4	BLINK!	5	5	4

Du scorte 22.5 av 25 mulige på temperaturvarslingen din!
Karakteren din blir dermed S .
Hirlam20 scorte 17 av 25 så du varslet bedre enn Hirlam20.
Hirlam10 scorte19.5 av25 så du varslet bedre enn Hirlam10.
EC scorte 20.5 av 25 , så du varslet bedre enn EC.

- Forecaster receives automatic email comparing own subjective T forecast with observations:

  $|\Delta T| \leq 1$ °C ➠ 1 point
  $|\Delta T| \leq 2$ °C ➠ 0,5 point
  $|\Delta T| > 2$°C ➠ 0 point

- A summary score is calculated: Ratio of total personal points to maximum possible

- Similarly for NWP forecasts

- Scores only available to individual forecasters * **NOT** * to the bosses!

Case, NORWAY *(Korsmo)*

Personal verification feedback by e-mail !

File Edit View Go Message Tools Window Help

Subject: Verifikasjon for 2004-05-23 12:00:00 **From:** helen.korsmo@met.no 05/23/2004 03:00 PM

Avistabell for 23/05-2004, kl. 12:00, skrevet av magnuso, parameter: T2m

STED	OBS	SUBJ:12+24	SUBJ.SCORE	HIRLAM20: 00+36	HIRLAM10: 00+36	ECMWF 12+48
Alta	7	7	BLINK!	5	7	7
Andøya	6	6	BLINK!	6	6	7
Bardufoss	6	7	BLINK!	8	8	9
Bodø	7	7	BLINK!	7	7	8
Brønnøysund	8	8	BLINK!	7	8	8
Glomfjord	6	8	Bra!	8	8	8
Hammerfest	5	4	BLINK!	4	5	4
Karasjok	6	5	BLINK!	4	4	5
Kautokeino	6	6	BLINK!	4	6	7
Kirkenes	3	3	BLINK!	5	4	2
Lakselv	7	6	BLINK!	4	5	6
Longyearbyen	2	2	BLINK!	0	1	2
Mosjøen	7	9	Bra!	9	8	9
Nord-Solver	9	8	BLINK!	7	6	8
Nordreisa	6	7	BLINK!	7	7	10
Rustefjelbma	4	4	BLINK!	5	4	4
Røst	7	7	BLINK!	6	6	7
Saltdal	7	10	Oups..	10	11	10
Sletnes fyr	3	4	BLINK!	5	5	4
Sortland	5	7	Bra!	7	7	7
Svolver	7	7	BLINK!	7	7	7
Torsvåg fyr	6	5	BLINK!	6	6	6
Tromsø	6	6	BLINK!	7	7	7
Vadsø	3	4	BLINK!	5	5	2
Vardø	4	4	BLINK!	5	5	4

Du scorte 22.5 av 25 mulige på temperaturvarslingen din!
Karakteren din blir dermed S .
Hirlam20 scorte 17 av 25 så du varslet bedre enn Hirlam20.
Hirlam10 scorte19.5 av25 så du varslet bedre enn Hirlam10.
EC scorte 20.5 av 25 , så du varslet bedre enn EC.

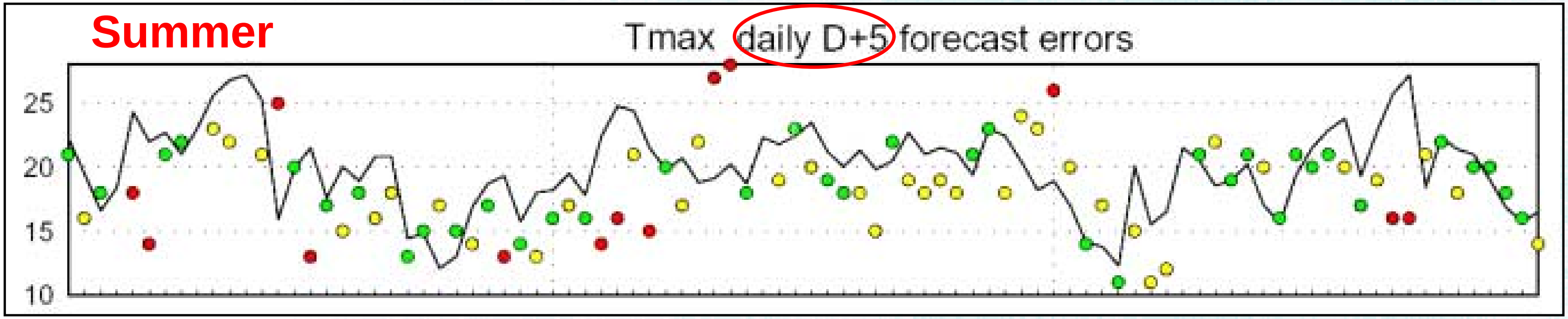
You scored 22.5 of 25

Hirlam20 scored 17.0 of 25

Hirlam10 scored 19.5 of 25

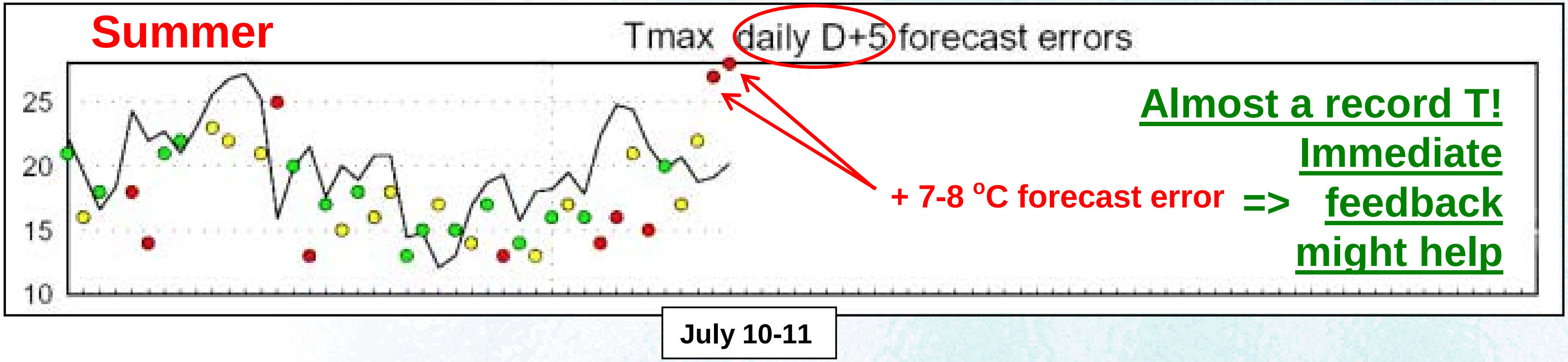
ECMWF scored 20.5 of 25

Case, FINLAND *(single site in Lapland up north)*



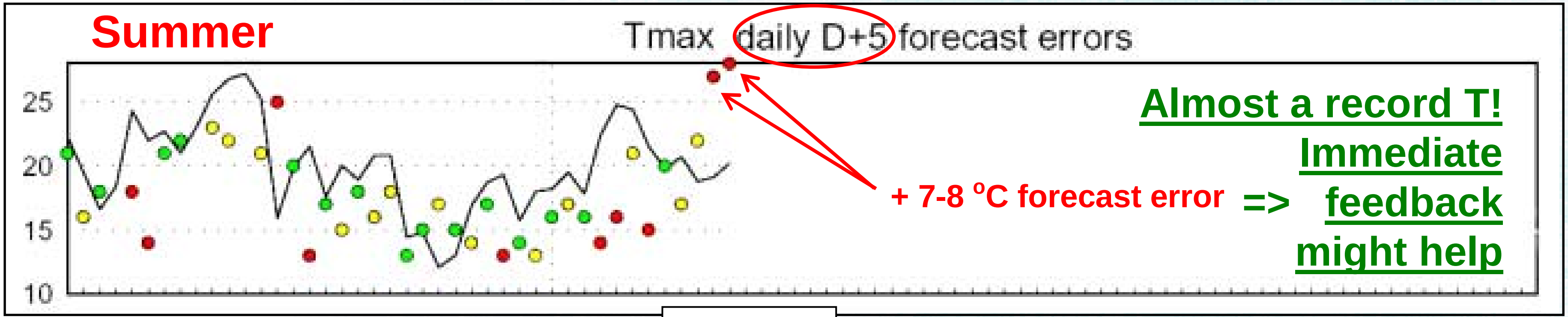
From: FMI operational verification system, **present**

Case, FINLAND *(single site in Lapland up north)*



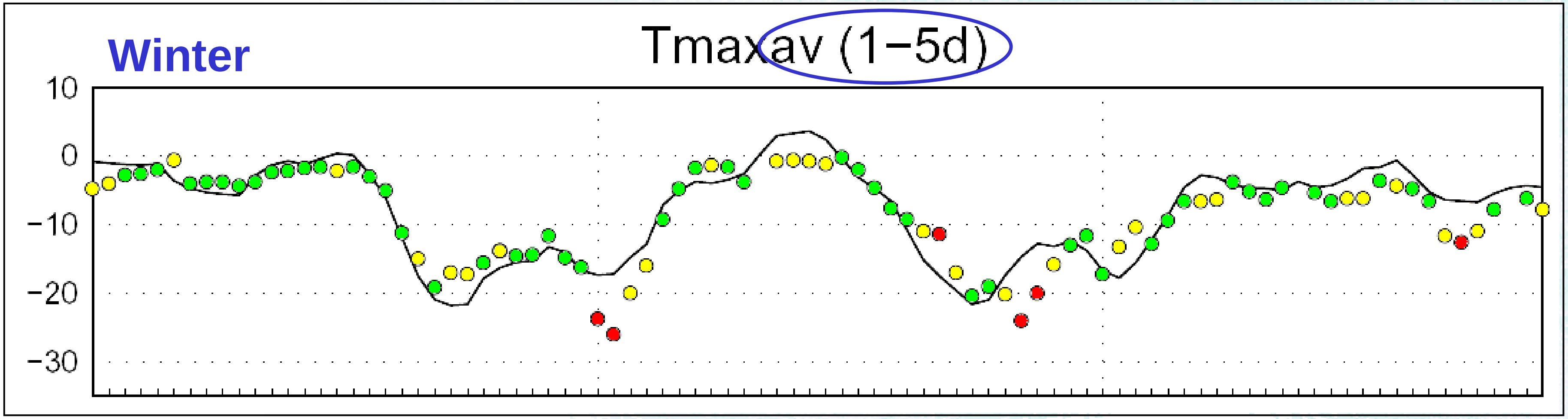
From: FMI operational verification system, present

Case, FINLAND *(single site in Lapland up north)*



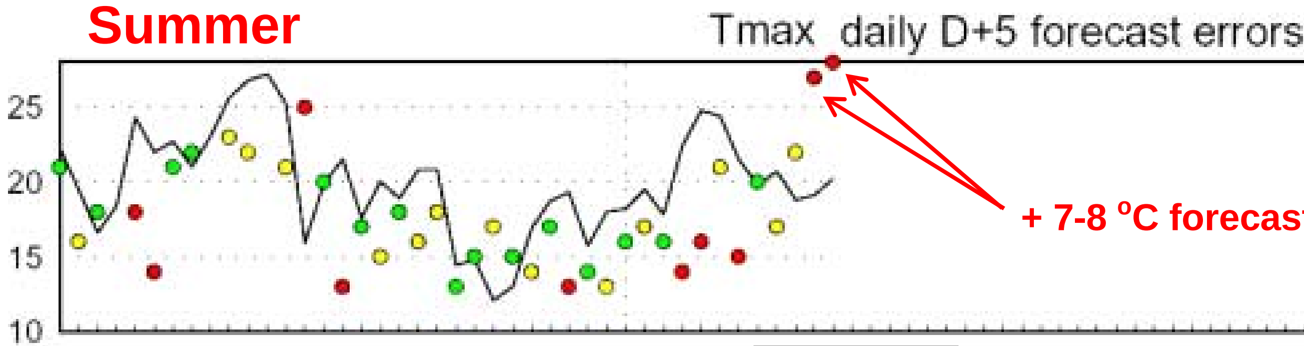
July 10-11

From: FMI operational verification system, present



Case, FINLAND *(single site in Lapland up north)*

Summer



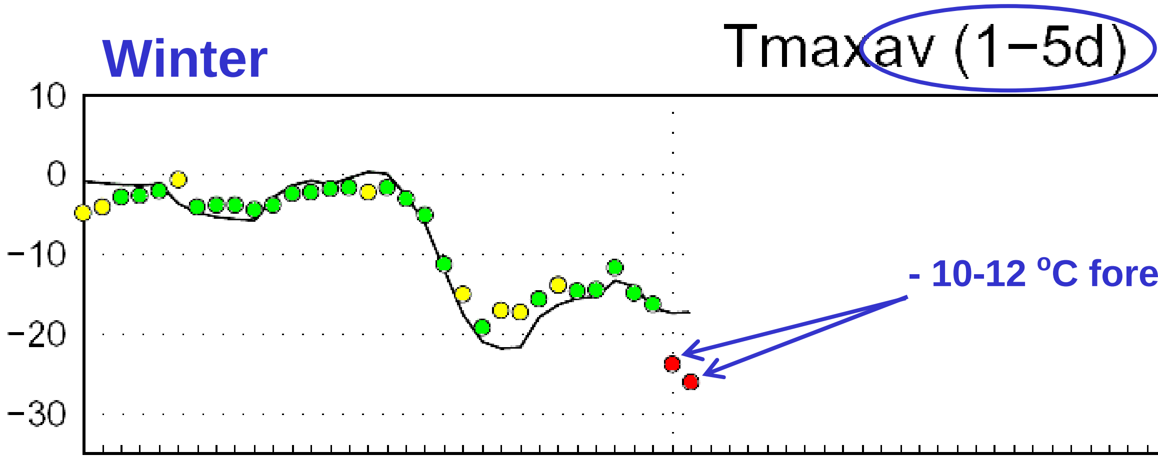
+ 7-8 ^oC forecast error **=>**

Almost a record T!
Immediate
feedback
might help

July 10-11

From: FMI operational verification system, present

Winter



- 10-12 ^oC forecast error **=>**

Fcr expects
"5th day" T
dropping
to - 50 ^oC
i.e. alltime record!

January 1-2

Outline:

- Introduction *…I guess you heard it already*
- General guidelines
- Verification measures for continuous variables
- Verification measures for categorical events
- Verification measures for probability forecasts
- Forecast value aspects
- Examples of practices at selected NWSs – scattered here and there during the presentation

First notes on (operational) verification:

- ✓ An essential daily, real-time, online practice in the operational forecasting environment
- ✓ A fundamental means to improve weather forecasts and services
- ✓ An act *(or even "art"?)* of countless methods and measures
- ✓ An active feedback and dialogue process between forecasters, modellers, developers, decision-makers, customers, public… thus serving the whole *(meteorological)* community

Principles of (all) verification:

- ✓ Verification activity has value only if the generated information leads to a decision about the forecast or the system being verified
  - User of the information must be identified
  - Purpose of the verification must be known
- ✓ No single verification measure can provide complete information about forecast quality
- ✓ Forecasts need be formulated in a verifiable form

=> e.g. How to verify worded forecasts?

Goals of (operational) verification:

✓ "Administrative"

- Feedback process to operational forecasters
- Monitor the quality of forecasts and potential trends in quality
- Feedback process to decision-makers, customers, public
  - Justify costs of provision of weather services
  - Justify additional or new equipment, models, …

✓ "Scientific"

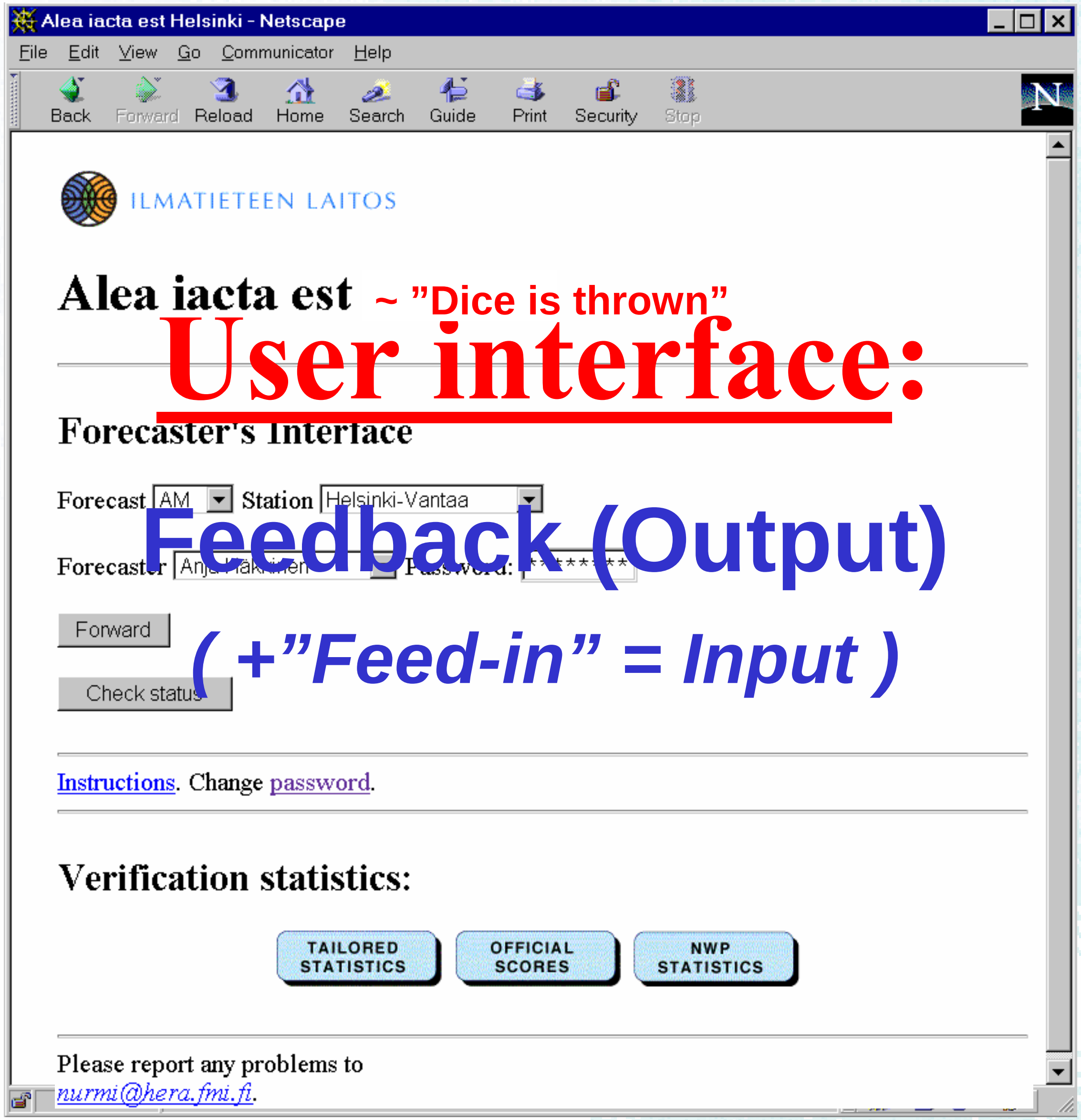
- Feedback process to modellers and developers
  - Identify strengths or weaknesses of a (NWP) forecast or guidance product leading to improvements, i.e. provide information to direct R&D

✓ "Value" *(not discussed here)*

<u>Operational online verification</u>, *“State-of-the-Art”*

- ✓ <u>Comprehensive evaluation</u> of <u>all</u> forecast(<u>er</u>)s
- ✓ Stratification <u>and</u> aggregation (pooling) of results
- ✓ Statistics of <u>guidance</u> forecasts *- e.g. NWP, MOS, competitor*
- ✓ <u>Instant</u> feedback to forecasters
- ✓ Statistics of <u>individual</u> forecasters – *e.g. personal biases*
- ✓ <u>Comprehensive</u> set of <u>tailored verification measures</u>
- ✓ <u>Simplified measures</u> for laymen
- ✓ <u>Continuity</u> into history
- ✓ Covers/Serves <u>all users/customers</u>
- ✓ Covers/Serves <u>official monitoring / target scores</u>
- ✓ Clear and functional <u>Web user-interface</u>; including <u>user-guide, tutorial, glossary, …</u>

Case, FINLAND *(very briefly)*



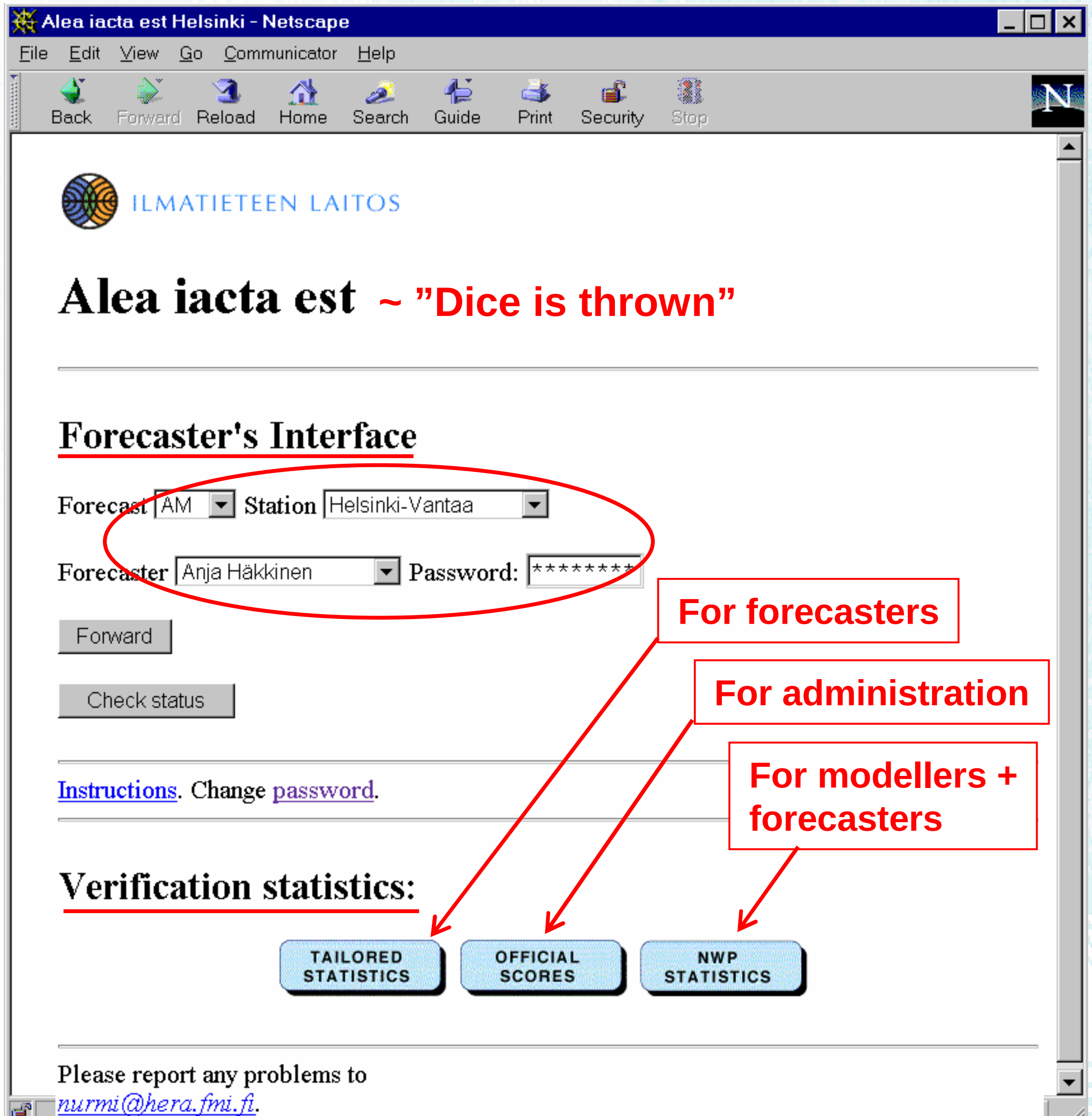
~ "Dice is thrown"

User interface:

Feedback (Output)

(+"Feed-in" = Input)

Case, FINLAND *(cont'd)*



Case, FINLAND *(cont'd)*

The integrated verification system *(somewhat outdated)*

- **W W W application**
  - Instant feedback of forecasts vs. obs
    - Timeseries graphs of forecasts vs. obs
    - "Black-listing" of missing forecasts
  - Time-averaged statistics
    - Error distributions
    - ME, MAE Skill Scores
    - Reliability diagrams ... etc.
- **Reporting Tool**
  - **25** years of CFO forecasts
  - RO forecasts since 1997
  - "All-known" verification measures
  - Annual, quarterly, monthly statistics of selected scores

Case, FINLAND *(cont'd)*

"Old" *(outdated)* Verification system: only 6 stations

Forecast input via:

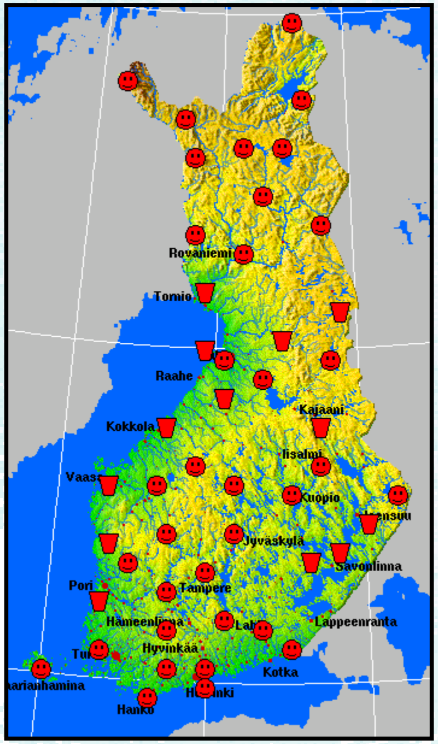
"intranet spreadsheets"

Under construction!

"New" Verification system: ~ 100 stations

Forecast input via: Grid Editor !

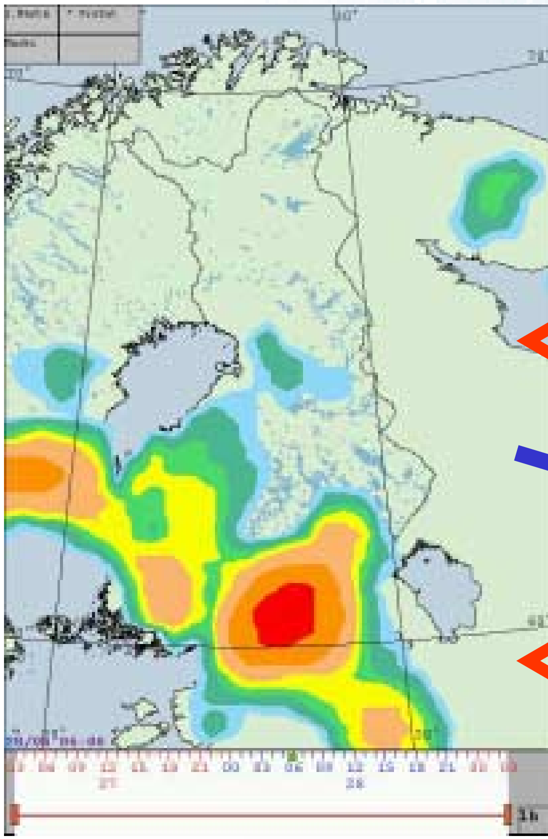
i.e. from points to grids



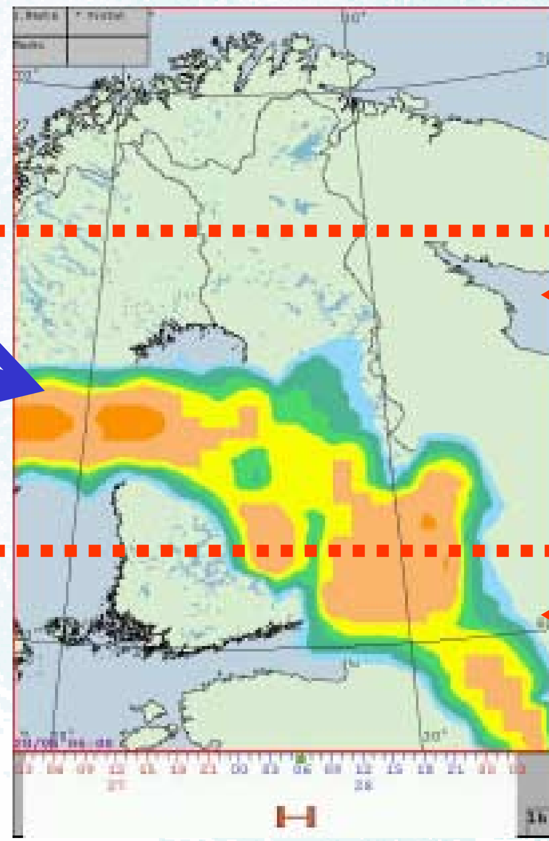
Case, FINLAND *(cont'd)*

The Forecasting-Verification Process *(idealistically)*

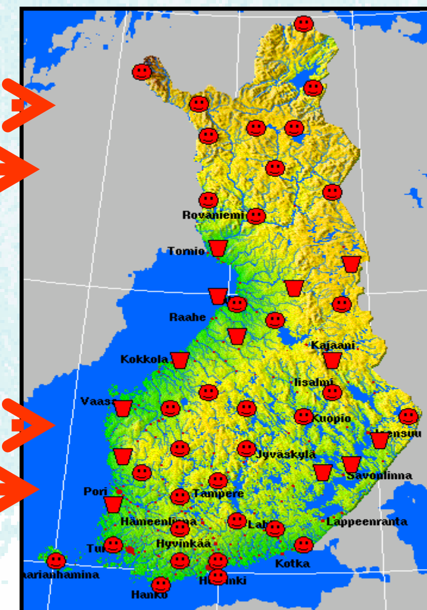
NWP guidance



Grid Edited End Product

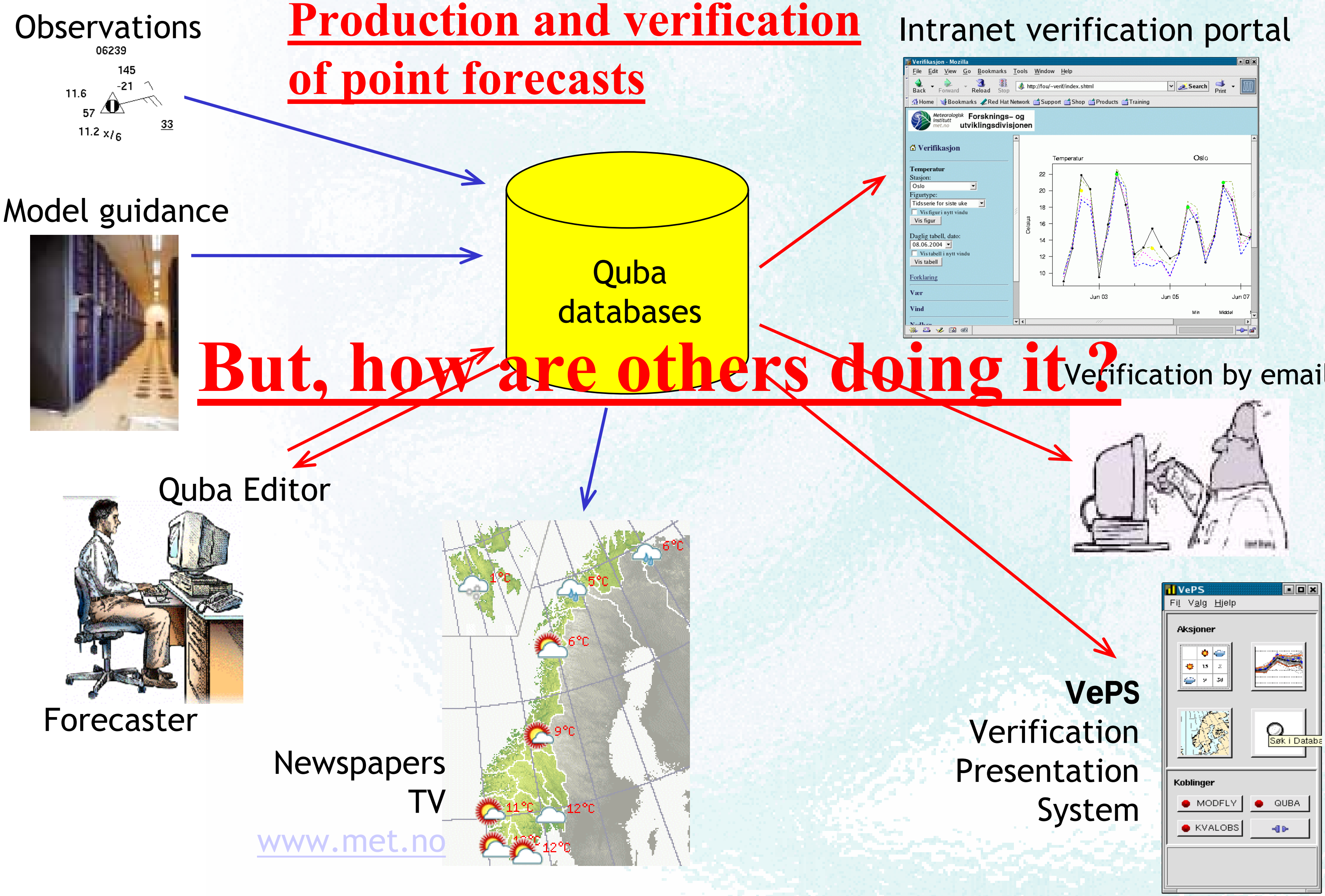


Verification



Case, NORWAY *(Korsmo)*

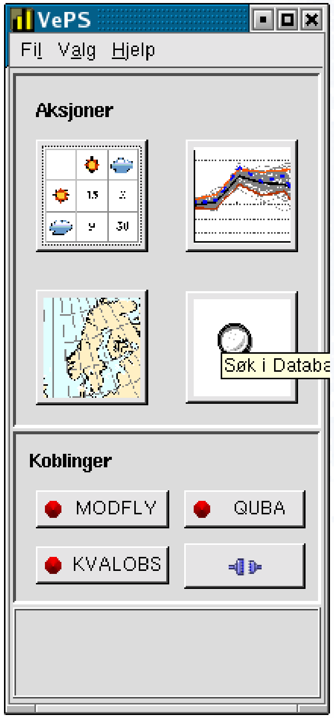
Production and verification of point forecasts



But, how are others doing it?

Case, NORWAY *(Korsmo)*

VePS – Verification Presentation System



- ✓ Graphical user interface to produce verification results on demand
- ✓ Forecasters log in to check their own verification results
- ✓ Timeseries, contingency tables, scatterplots, geographical distributions produced according to user specifications
- ✓ Forecasts and observations read from the Quba historical database in real time
- ✓ Connected to the meteorological workstation Diana, where the verification results are displayed on the map

Case, SLOVENIA *(Razinger)*

Verification score

1) *select region for verification analysis*

- Area: asi
- Country: Czech Republic
- Borders: latitude Nord / longitude West / longitude East / latitude South
- Station: LJUBLJANA/BEZIGRAD (t)

2) *select time range*

- from: 1 . 9 . 2002 to: 19 . 11 . 2003
- from: Sep . 2002 to: Nov . 2003
- last 10 days / last month / last 3 months / all data range

3) *select at least one model*

AS00 AS12 AST00 AST12 AL00 AL12 EC12 ETA12

4) *select variable for verification*

surface variables:

T2m	T2m min	T2m max	
T2m corr.	T2m min corr.	T2m max corr.	10m-FX
10m-FF	10m-DD	10m-U	10m-V
Pmsl	RH2m	CC	RR24h

pressure level: 925 850 700 500 250

variable: H T RH FF DD U V

Continue>>

CC-cloud cover, RR-precipitation, RH-relative humidity, FF-wind velocity, DD-wind direction, U-zonal wind component, V-meridional wind component, FX-wind gusts

Verification score 2/2

Selected region:	SLOVENIA
Selected period:	period between 2002-9-1 and 2003-11-20
Selected models:	as12 ast12
Selected variable:	T2m

5) *select verification score*

ME MAE RMSE SD

6) *select data filtering conditions*

include only stations above m and below m

reject data if difference is more than 0 °C

7) *and finally, select data view*

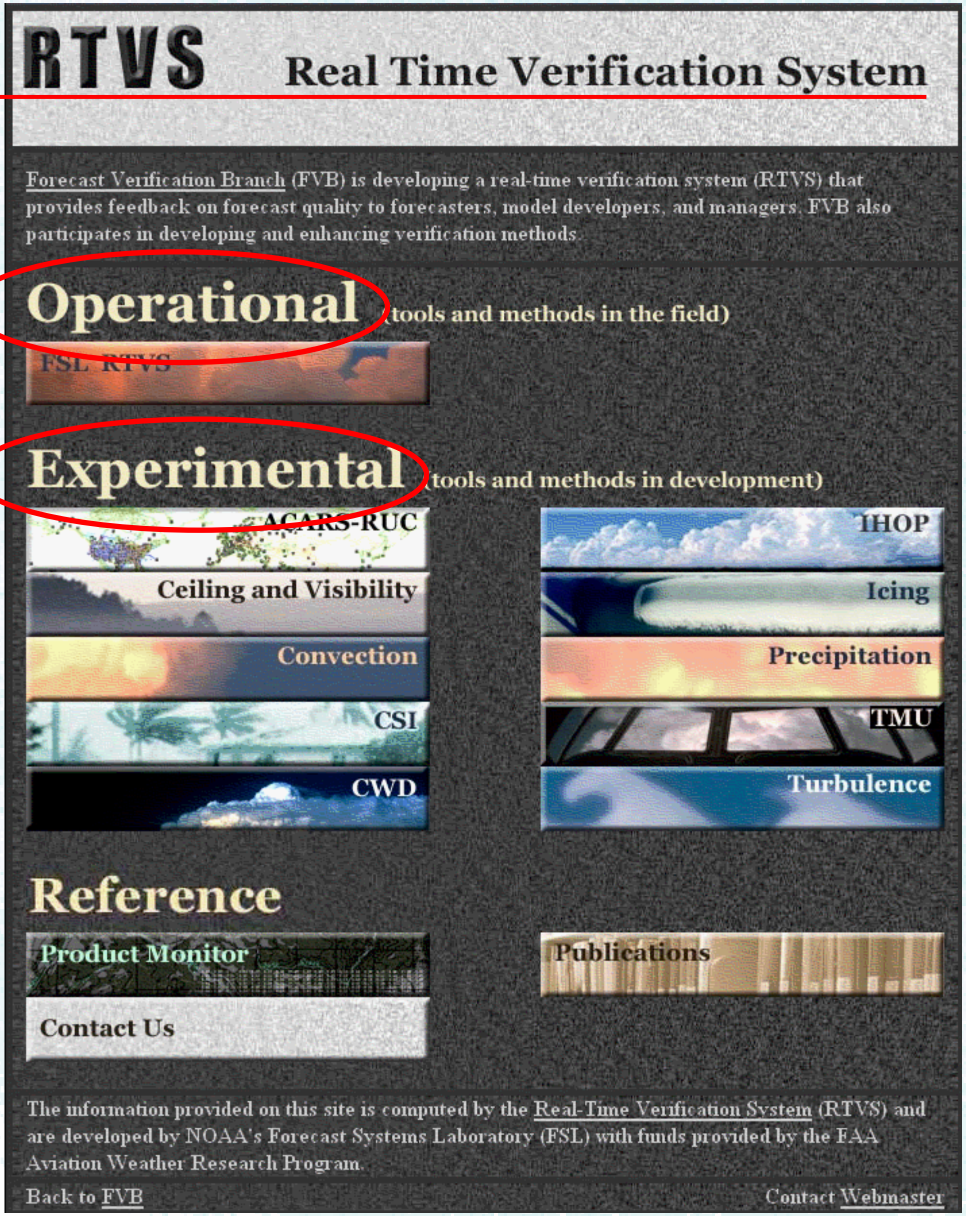
- scatterplot FC all — graph
- number/FC — table
- score/FC — map FC 00
- score/time HH 00

Show scores>>

Separate from forecast production, i.e. input system ???

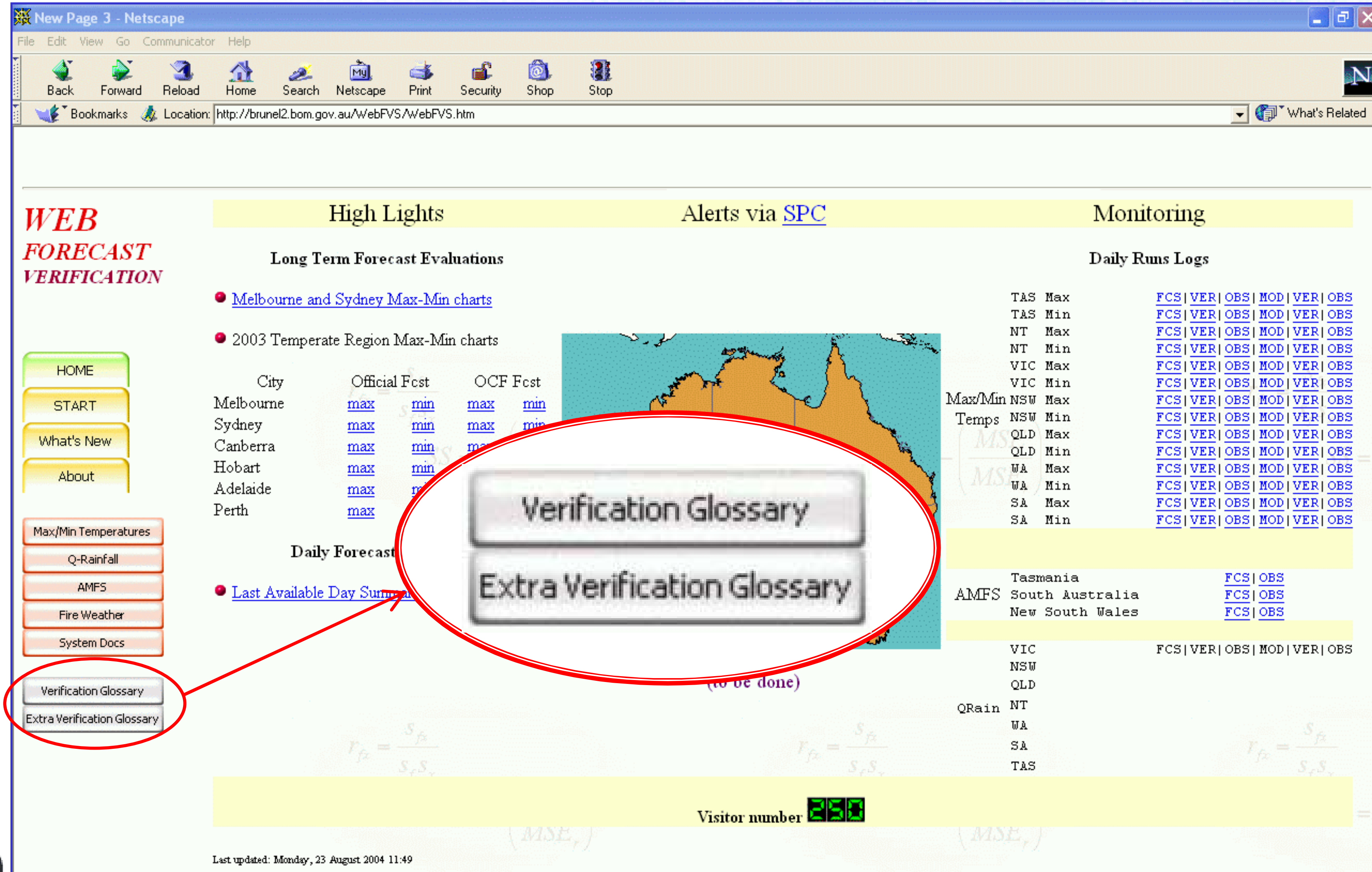
Case, USA *(Loughe et al.)*

- ✓ Developed by NOAA's Forecast Systems Laboratory (FSL)
- ✓ Cornerstone for verification within the FAA **Aviation Weather** Research Program (AWRP)
- ✓ Feedback to forecasters, model developers, managers



Case, Australia *(Ebert et al.)*

New Web Forecast Verification System (WebFVS)



Case, USA *(Brill; www.hpc.ncep.noaa.gov/npvu/qpfv)*

***** WARNING *****

The inherent quality of the verification statistics contains seasonal and geographical variations. Specifically, the scores contain degradation in…

(1) Northern latitudes during winter due to poor ground truth data when precipitation is frozen, (2) For the three RFC areas west of the Continental Divide due to an inconsistency in the sc… RFC areas just east of the Continental Divide (MBRFC … of poor ground truth data in the Rocky Mountain f… e latter two problems are being addressed.

… might a.o. cover information on potential misuse / misunderstanding of verification…

NWP verification *(Some of the points apply to all verification)*

- ”Easy” to implement *(in the free atmosphere)*
- Common, well-established practices, measures and scores exist
- Verification vs. numerical analysis & observations
- Comparison of model versions & different models
- Global, hemispheric, local areas, station-based (stratification)
- Stratification by forecast range, time, area, pressure level, …

Let’s have a look at NWP verification

- Geopotential, temperature, humidity, wind, …
- ME, MAE, RMSE, *(anomaly, tendency)* correlations, skill, …

(Some of the points apply to all verification)

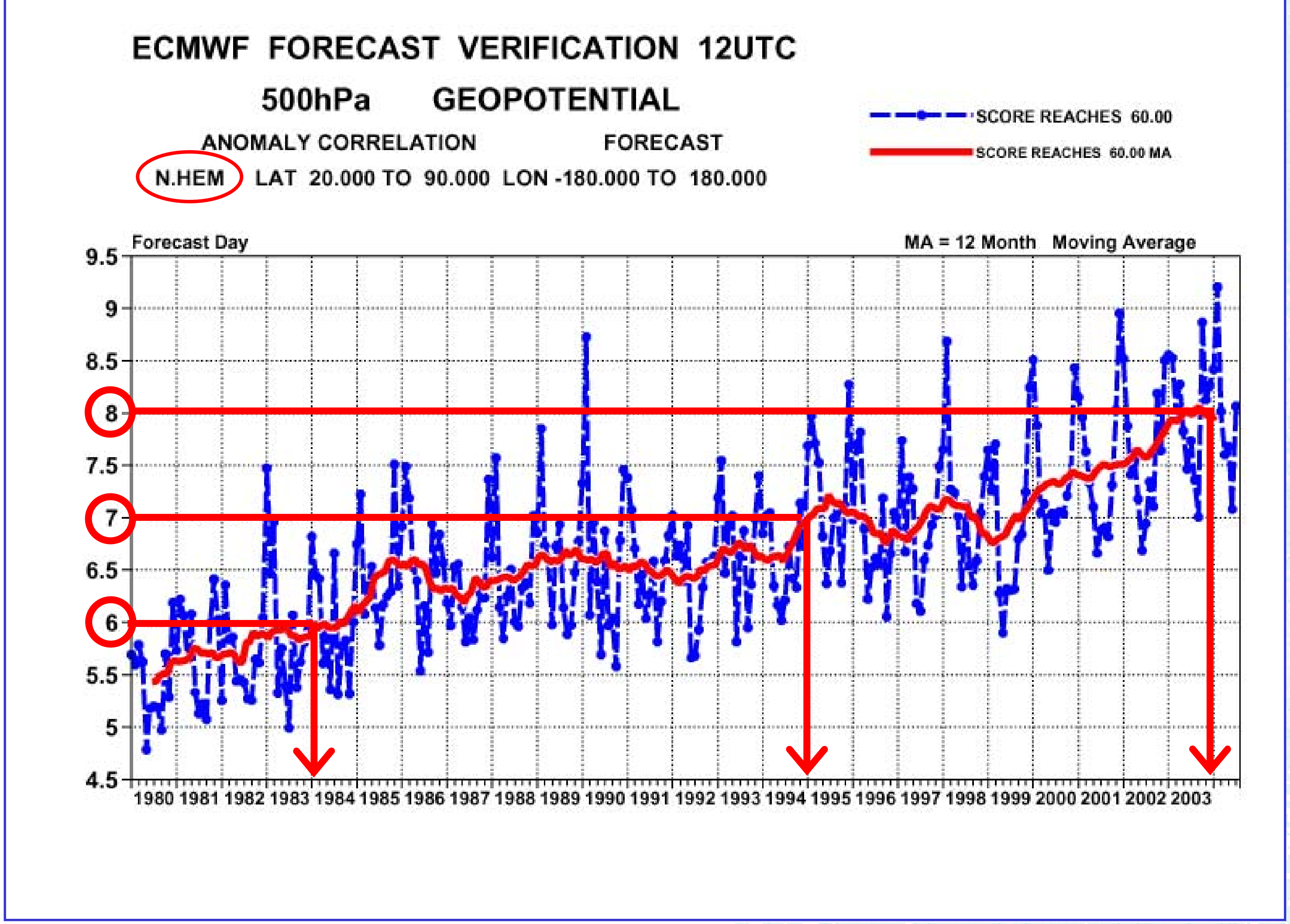
- Surface weather elements (T_{2m}, Td_{2m}, V10, R, N): comparison with MOS, PP, EP, …
- ”Special treatment” -> EPS
- **BUT**: Do the forecasters generally have easy, direct, automated online access to all of these ?

I doubt it, although they should have !

Case, ECMWF *(web)*

NWP

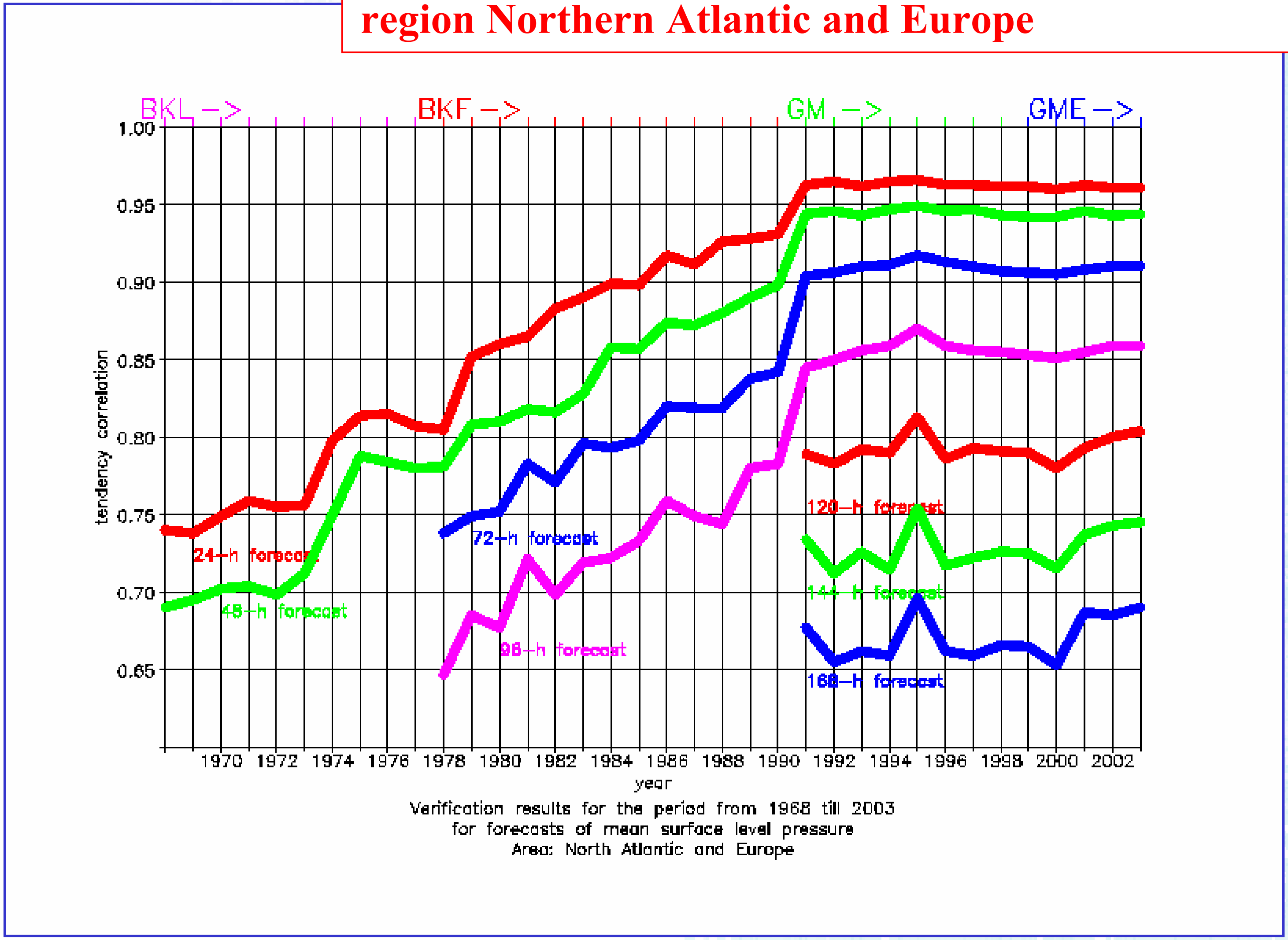
Official target: Extension of predictability, i.e. anomaly correlation skill > 60 %, at the rate of day per decade



Case, Germany *(Damrath)*

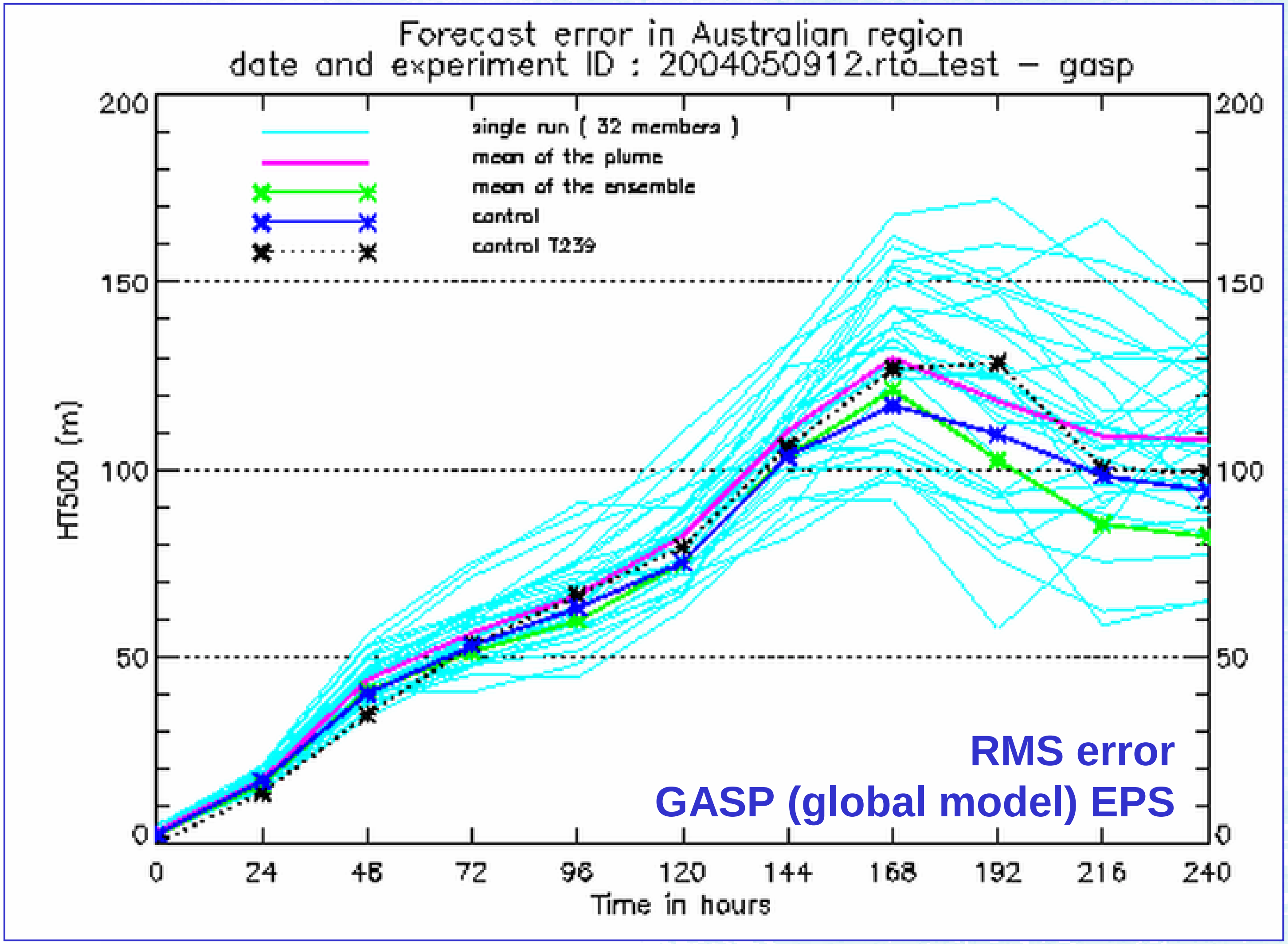
NWP

Time series of MSLP tendency correlation coefficient region Northern Atlantic and Europe



Case, Australia *(Ebert et al.)*

EPS verification [1]

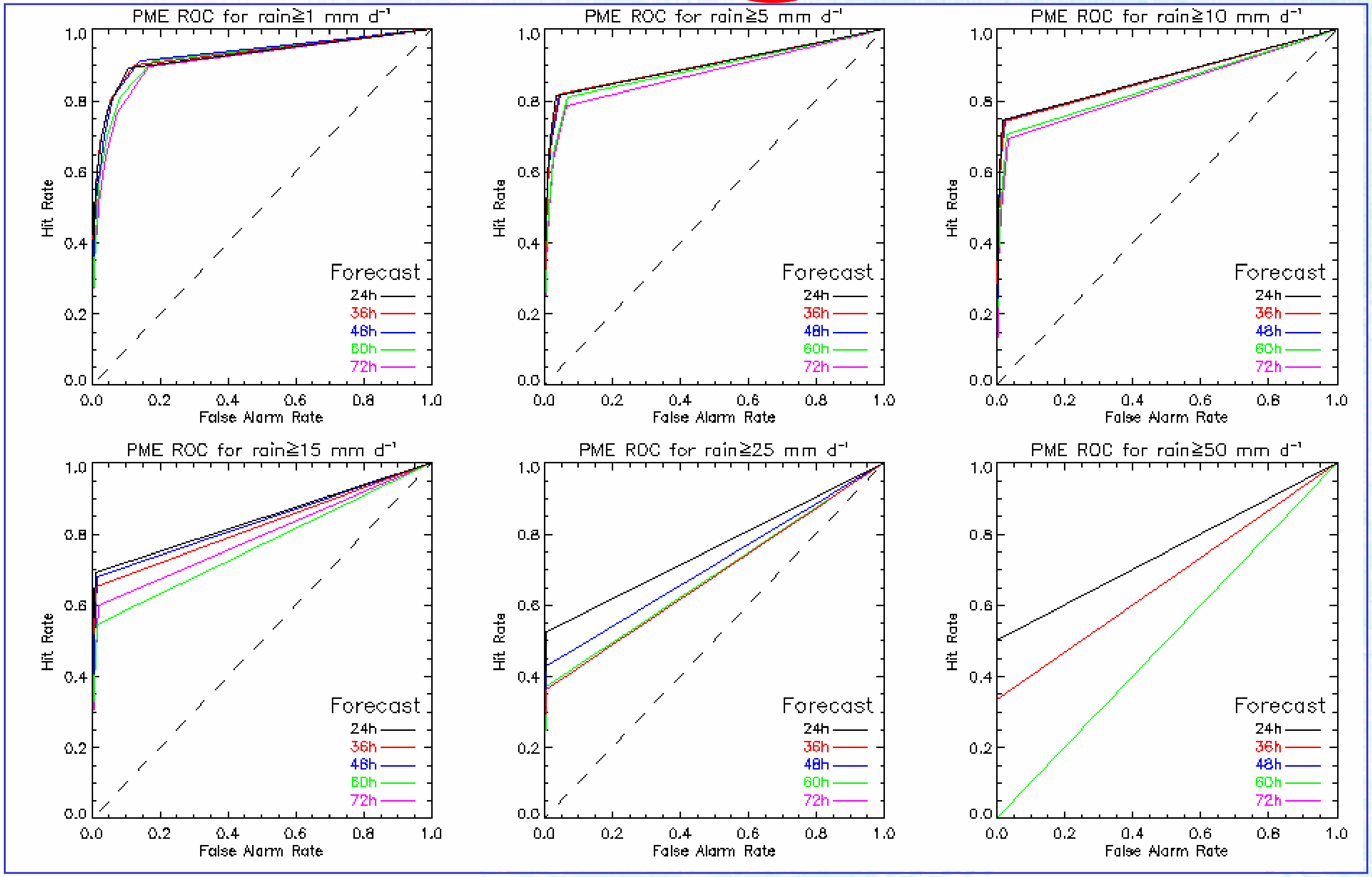


Case, Australia *(Ebert et al.)*

Applicable to all probabilistic, and even categorical, forecasts!

EPS verification 2

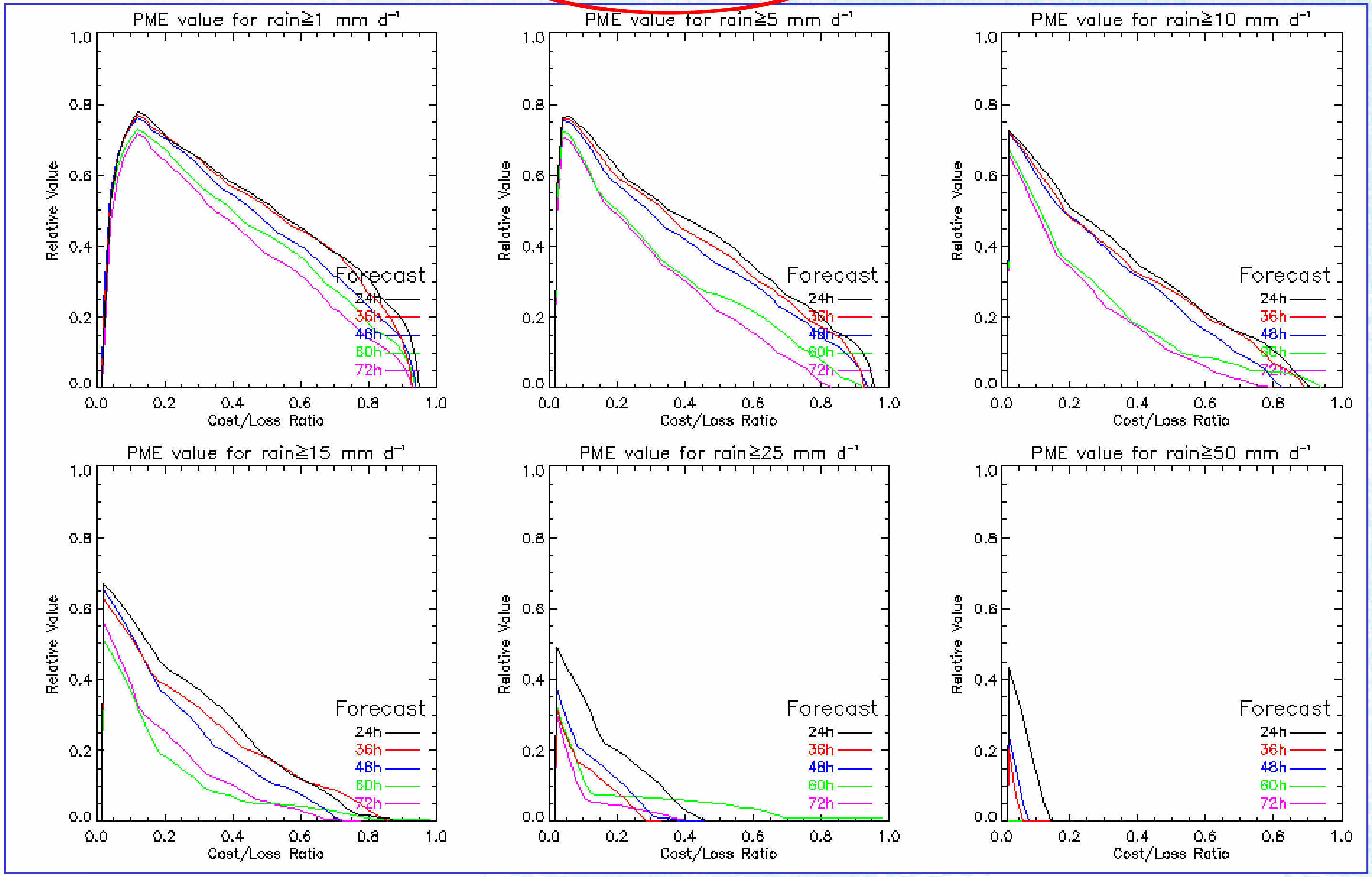
ROC – Poor Man's Ensemble



Case, Australia *(Ebert et al.)*

EPS verification 3

Relative value – Poor Man's Ensemble



Uncertainty - Why Probability Forecasts ?

- "… the widespread practice of ignoring uncertainty when formulating and communicating forecasts represents an extreme form of inconsistency and generally results in the largest possible reductions in quality and value."

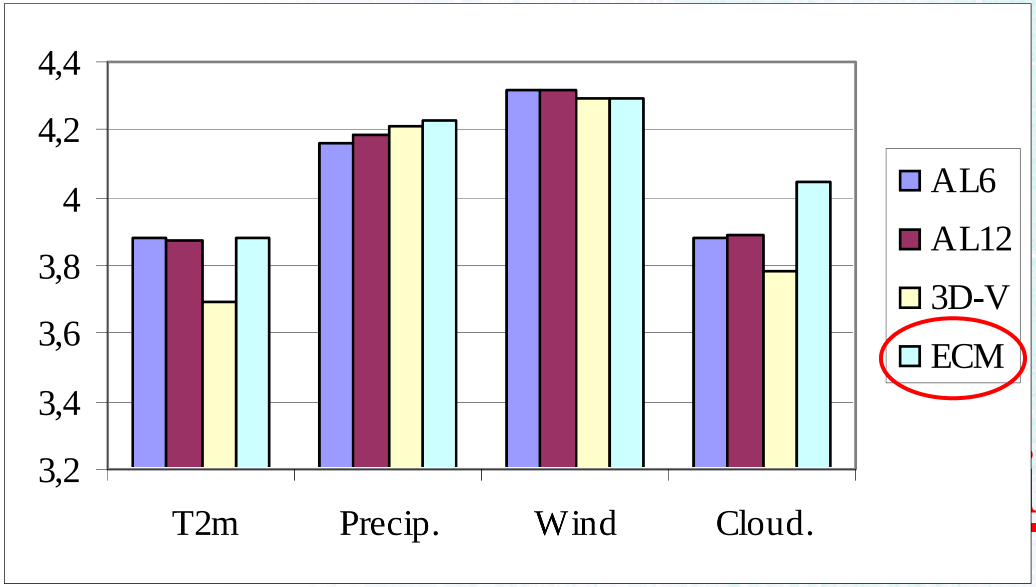
- Allan Murphy (1993)

Uncertainty - Why Probability Forecasts ?

©"… Go look at the weather,
I believe it's gonna rain"

- Legendary Chicago Blues Artist
Muddy Waters (early 1960s)
"Clouds in My Heart"

Case, Hungary *(Toth)*

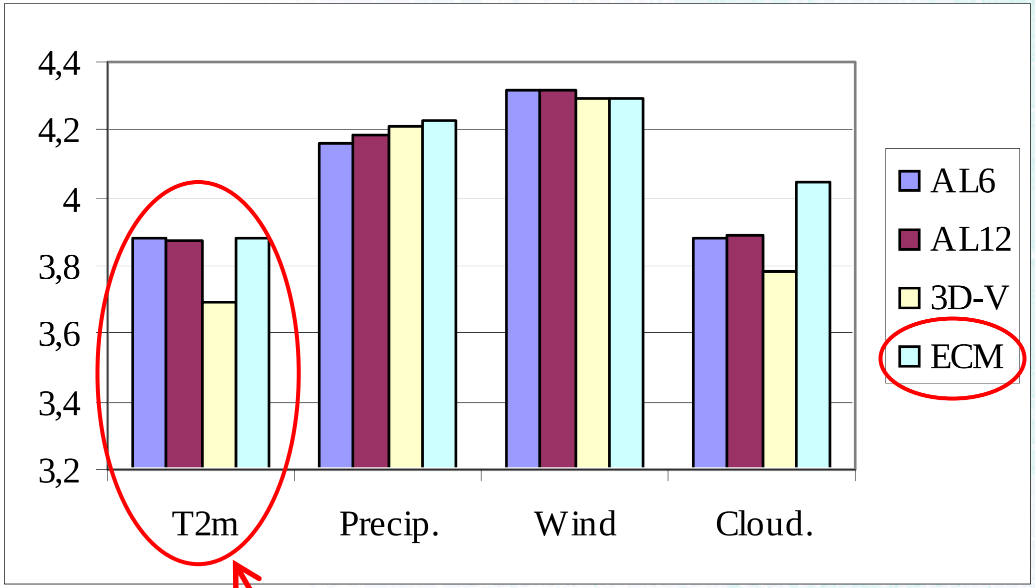


4 model versions subjectively evaluated

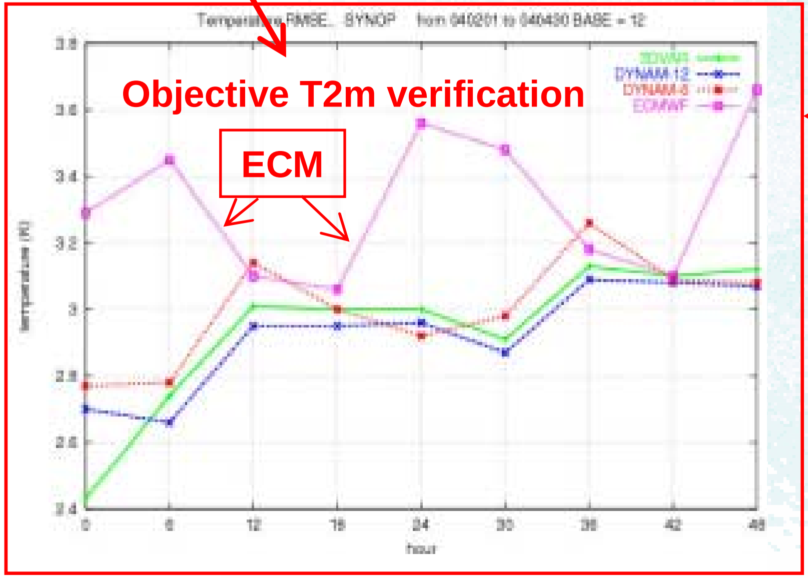
ication

(to be built in operational verification systems, e.g. to help choose the right model):

Case, Hungary *(Toth)*



4 model versions <u>subjectively</u> evaluated



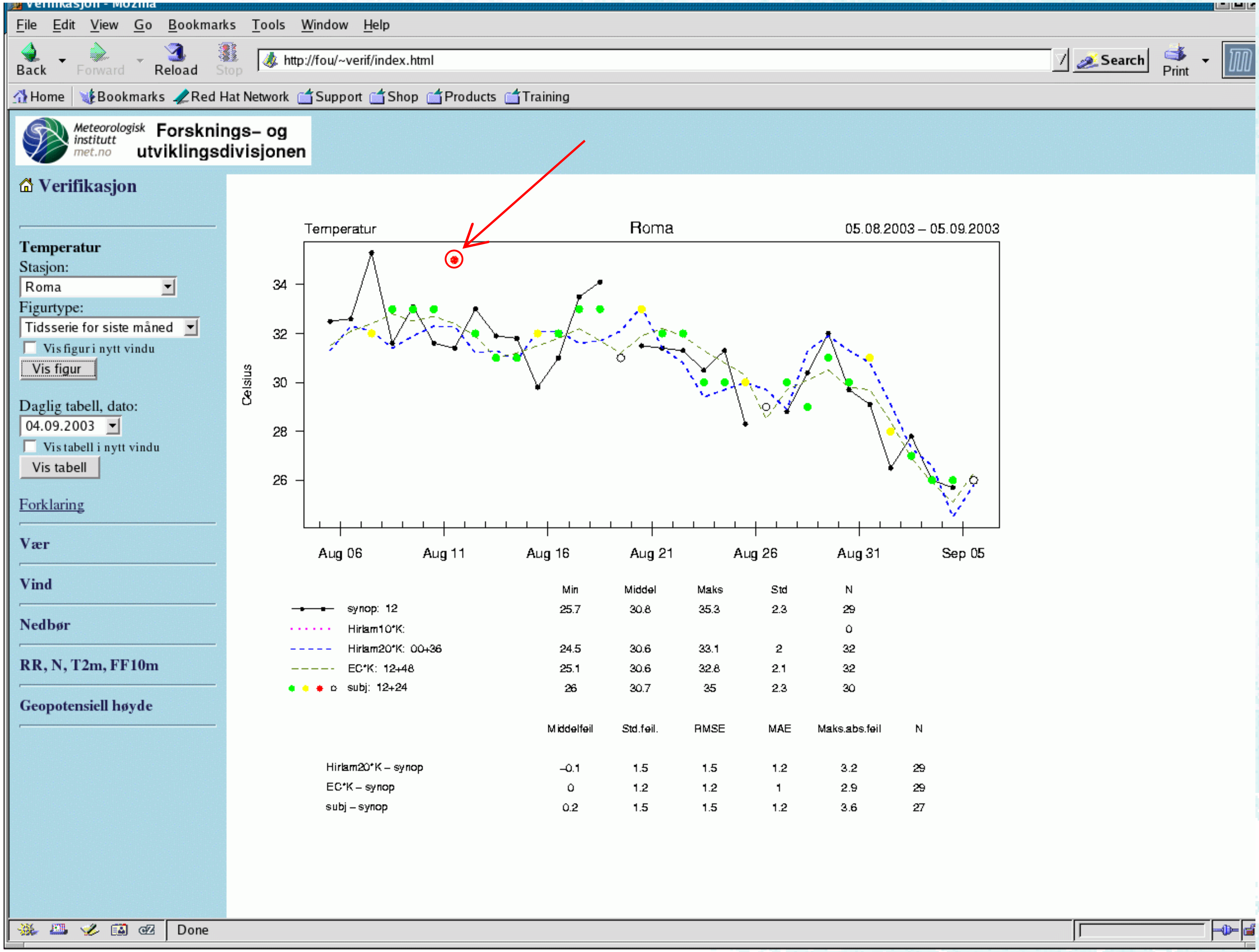
…compared with <u>objective</u> verification scores…

End product, public weather and rare event verification:

- Verification feedback in the **inter**net, rather than **intra**net, which often seems to be the case

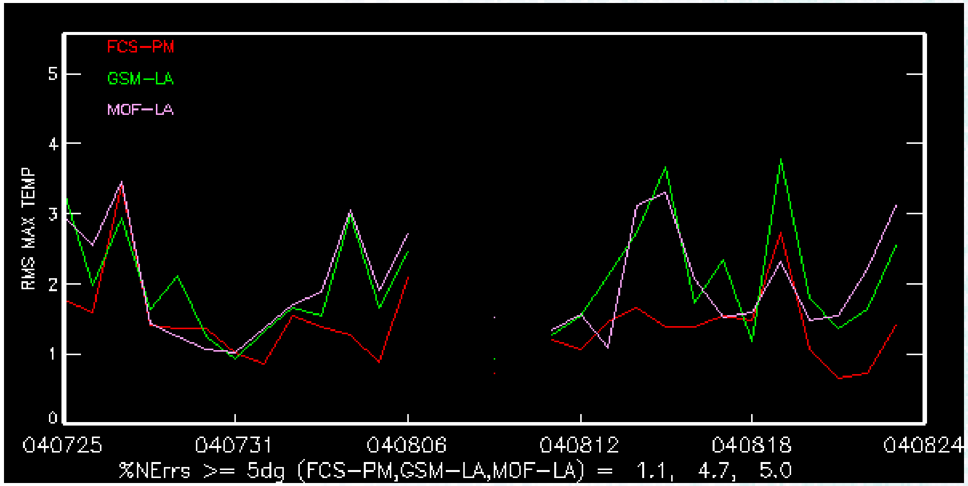
Case, NORWAY *(Korsmo)*

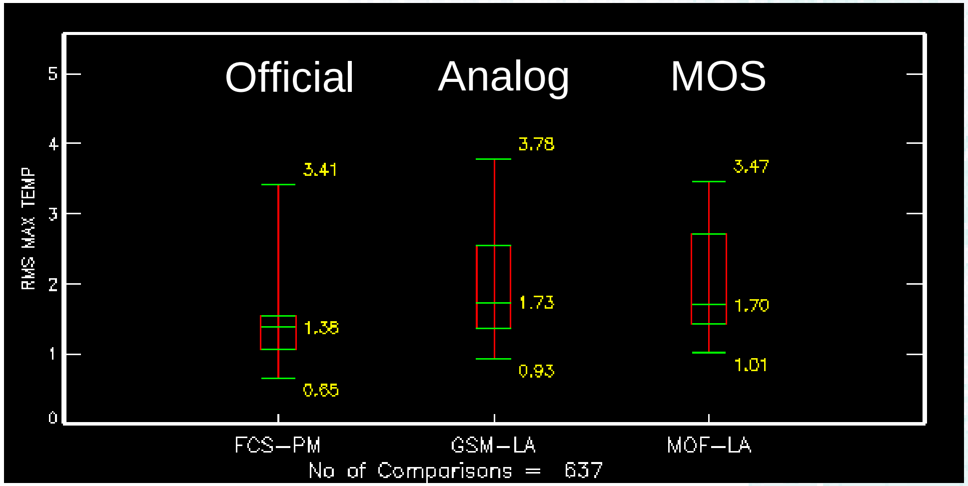
Intranet verification portal, example temperature time-series



Case, AUSTRALIA *(Ebert et al.)*

Official forecasts and objective (statistical) guidance at sites



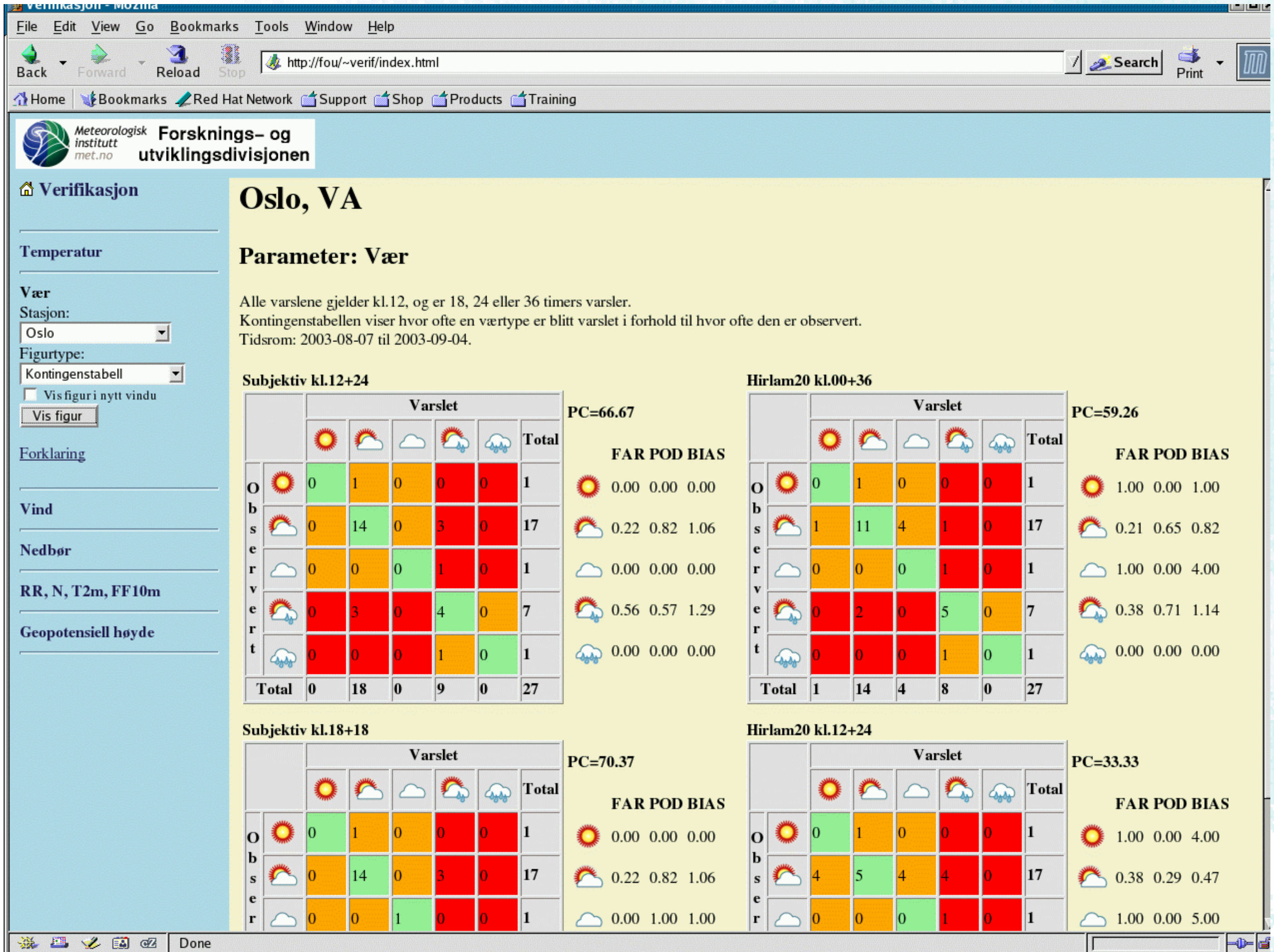


Site-based verification for most recent 30-day period

- ✓ **D+1 official (person) forecasts & two statistical forecasts**
- ✓ **T_{max} & T_{min}**
- ✓ **State-based verification statistics; large number of sites**
  - RMS error
  - Box plots
  - Number of errors > 4.5 ˚C

Case, NORWAY *(Korsmo)*

Intranet verification portal, example weather parameter



Oslo, VA

Parameter: Vær

Alle varslene gjelder kl.12, og er 18, 24 eller 36 timers varsler.
Kontingenstabellen viser hvor ofte en værtype er blitt varslet i forhold til hvor ofte den er observert.
Tidsrom: 2003-08-07 til 2003-09-04.

Subjektiv kl.12+24 — PC=66.67

Observert \ Varslet						Total
	0	1	0	0	0	1
	0	14	0	3	0	17
	0	0	0	1	0	1
	0	3	0	4	0	7
	0	0	0	1	0	1
Total	0	18	0	9	0	27

	FAR	POD	BIAS
	0.00	0.00	0.00
	0.22	0.82	1.06
	0.00	0.00	0.00
	0.56	0.57	1.29
	0.00	0.00	0.00

Hirlam20 kl.00+36 — PC=59.26

Observert \ Varslet						Total
	0	1	0	0	0	1
	1	11	4	1	0	17
	0	0	0	1	0	1
	0	2	0	5	0	7
	0	0	0	1	0	1
Total	1	14	4	8	0	27

	FAR	POD	BIAS
	1.00	0.00	1.00
	0.21	0.65	0.82
	1.00	0.00	4.00
	0.38	0.71	1.14
	0.00	0.00	0.00

Subjektiv kl.18+18 — PC=70.37

Observert \ Varslet						Total
	0	1	0	0	0	1
	0	14	0	3	0	17
	0	0	1	0	0	1

	FAR	POD	BIAS
	0.00	0.00	0.00
	0.22	0.82	1.06
	0.00	1.00	1.00

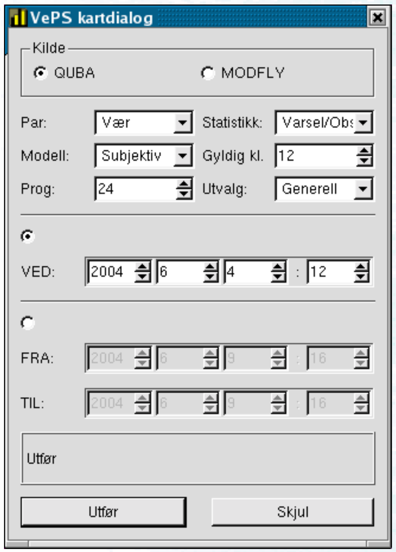
Hirlam20 kl.12+24 — PC=33.33

Observert \ Varslet						Total
	0	1	0	0	0	1
	4	5	4	4	0	17
	0	0	0	1	0	1

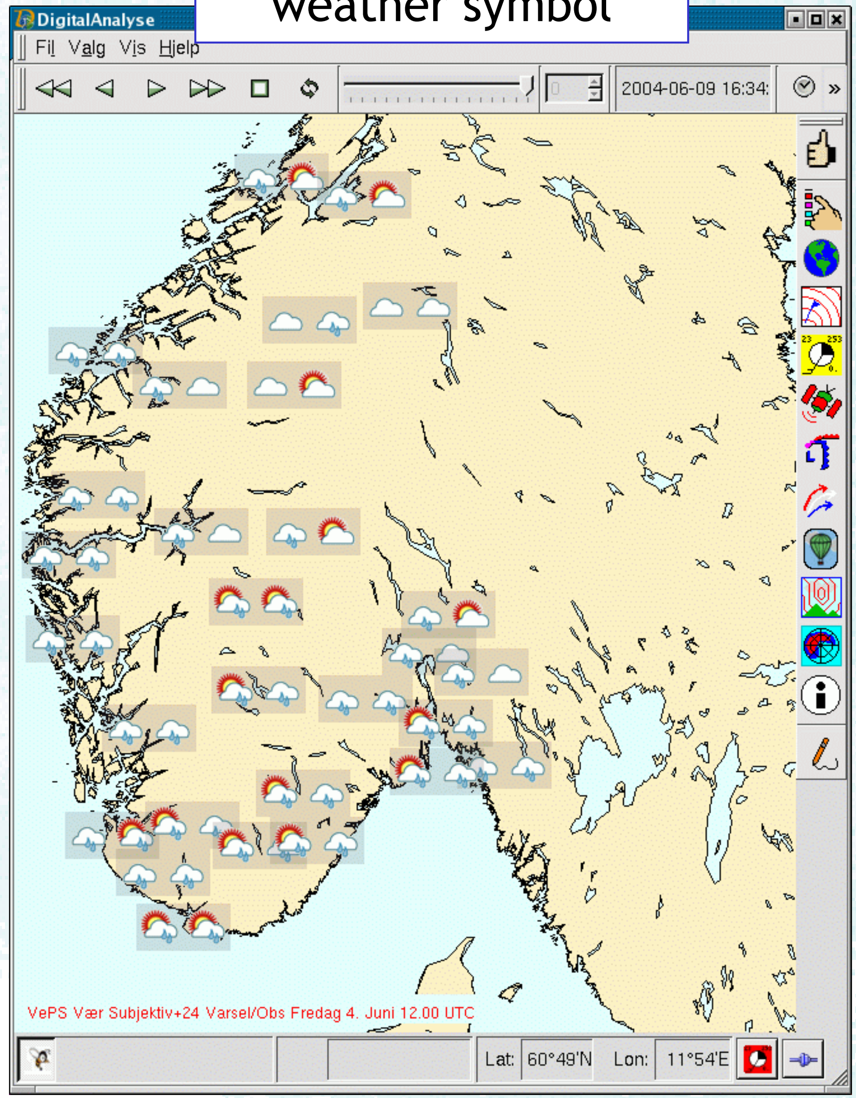
	FAR	POD	BIAS
	1.00	0.00	4.00
	0.38	0.29	0.47
	1.00	0.00	5.00

Case, NORWAY *(Korsmo)*

Intranet verification portal, example map plotting

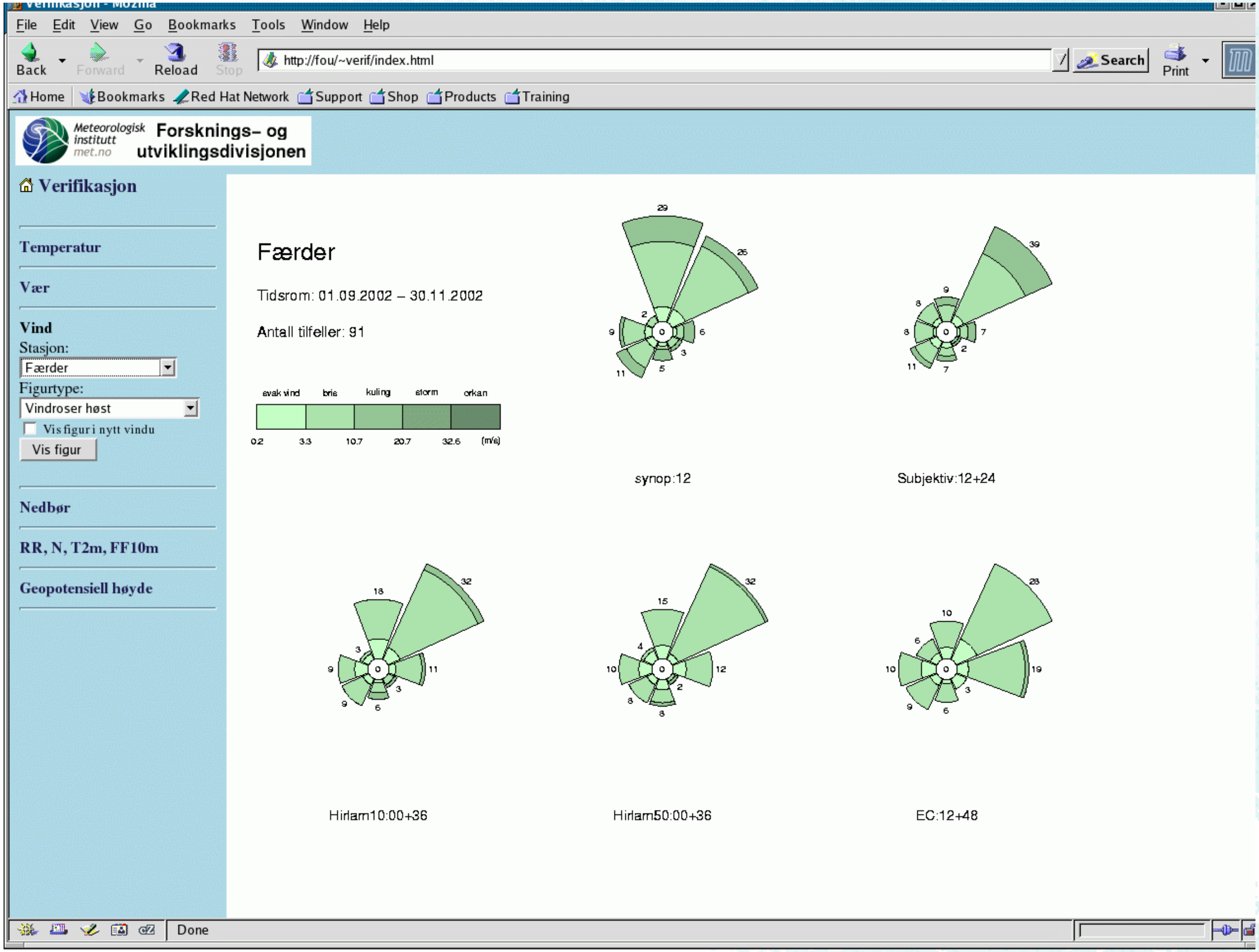


Forecast / Observed weather symbol



Case, NORWAY *(Korsmo)*

Intranet verification portal, example wind roses



Case, FINLAND *(c Mikael Frisk)*

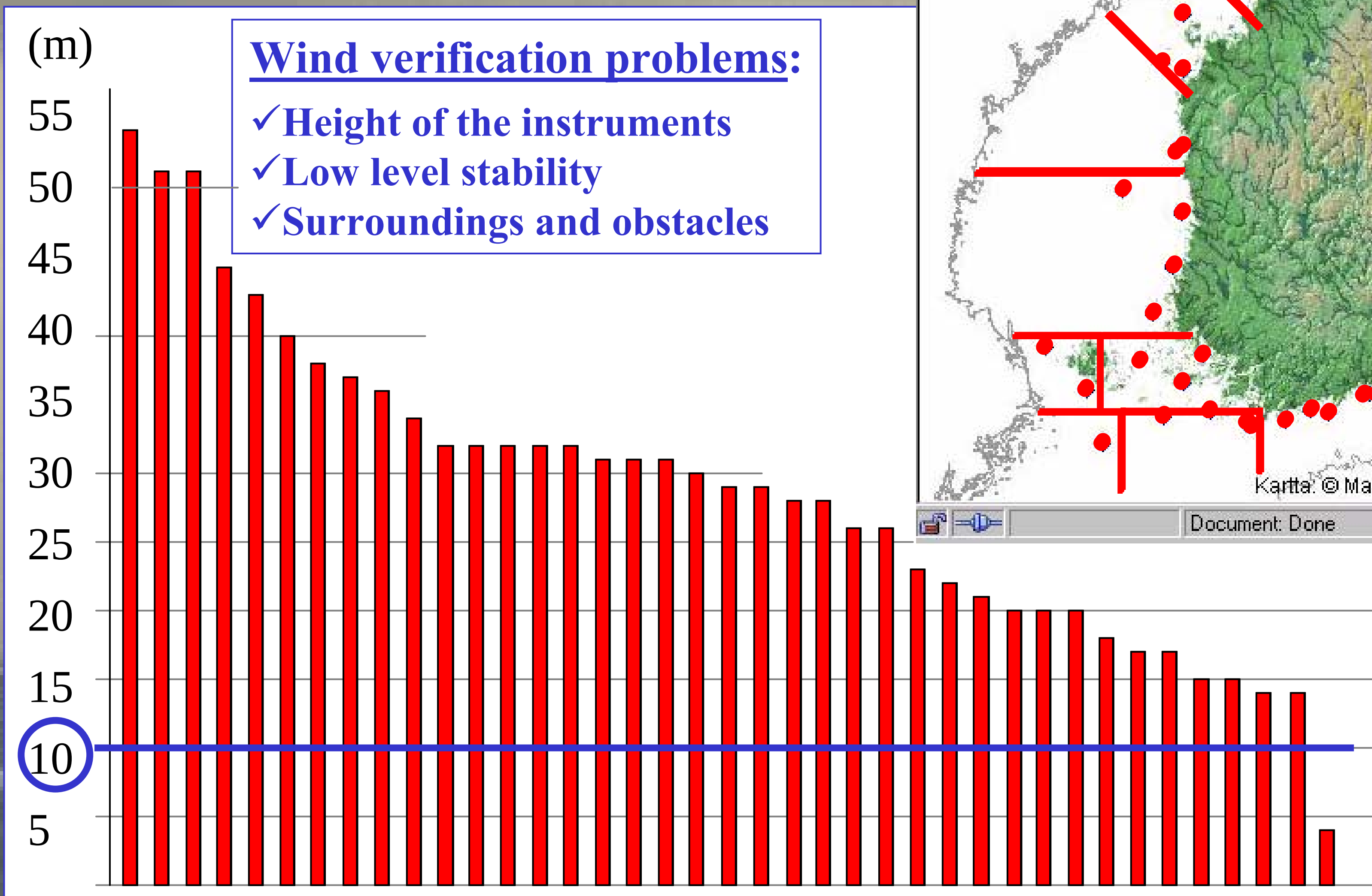
…even we have rare events & severe weather

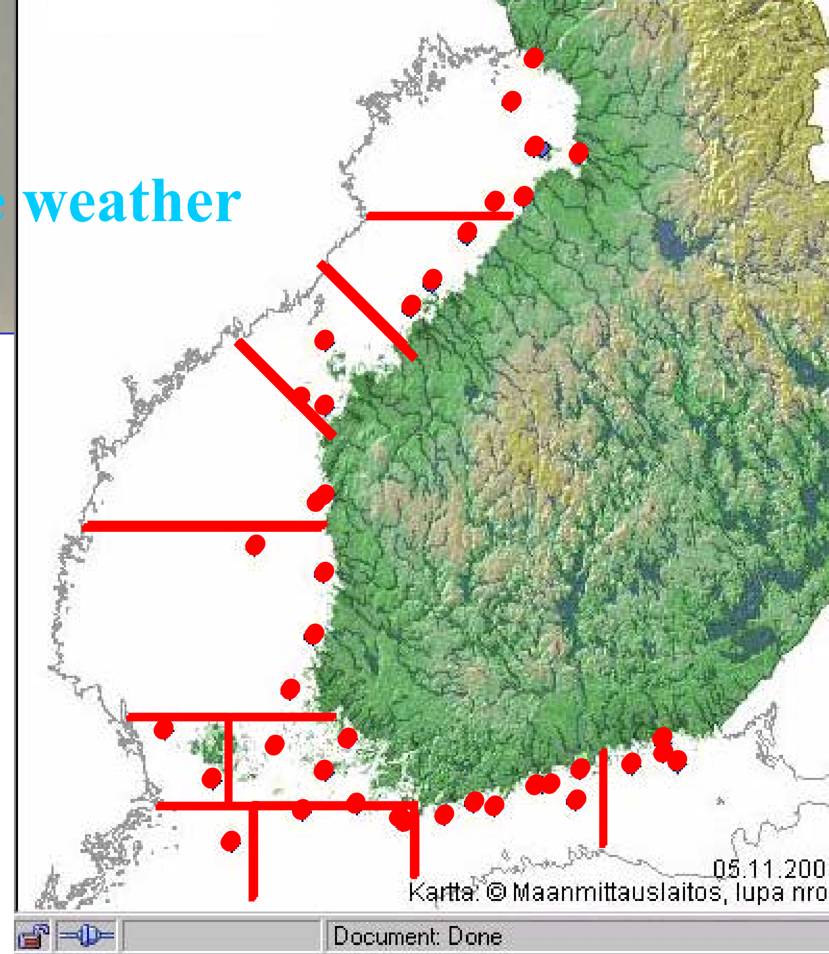
Case, FINLAND

…even we have rare events & severe weather

Wind verification problems:

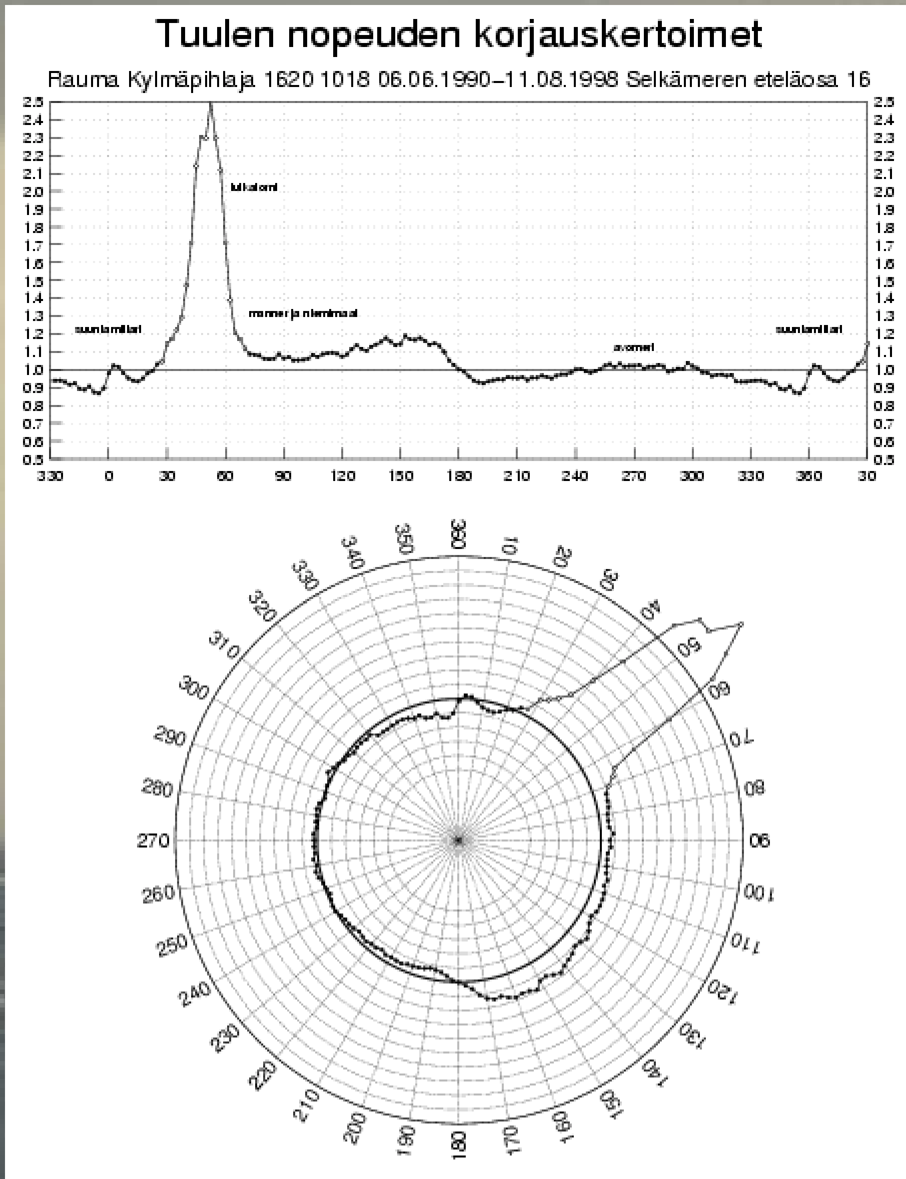
- ✓Height of the instruments
- ✓Low level stability
- ✓Surroundings and obstacles

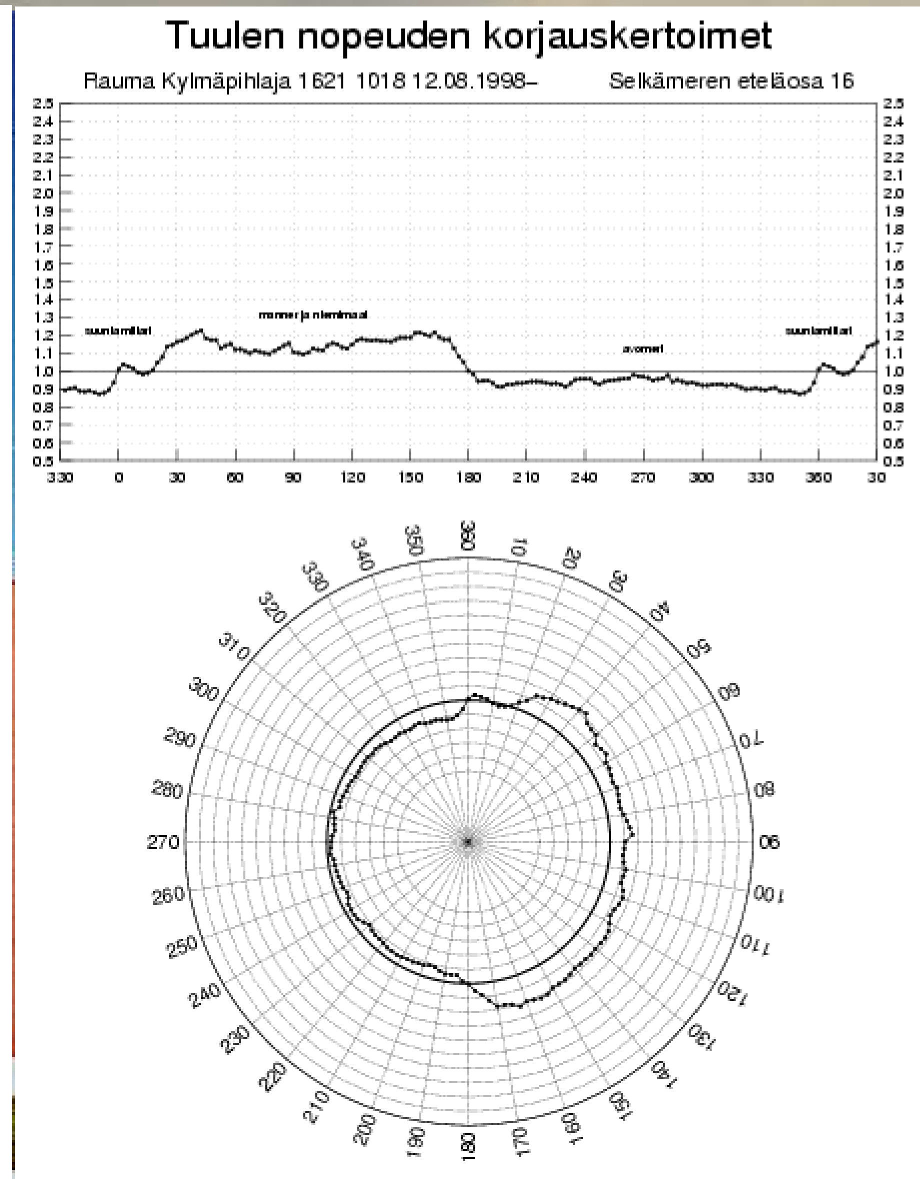




Case, FINLAND

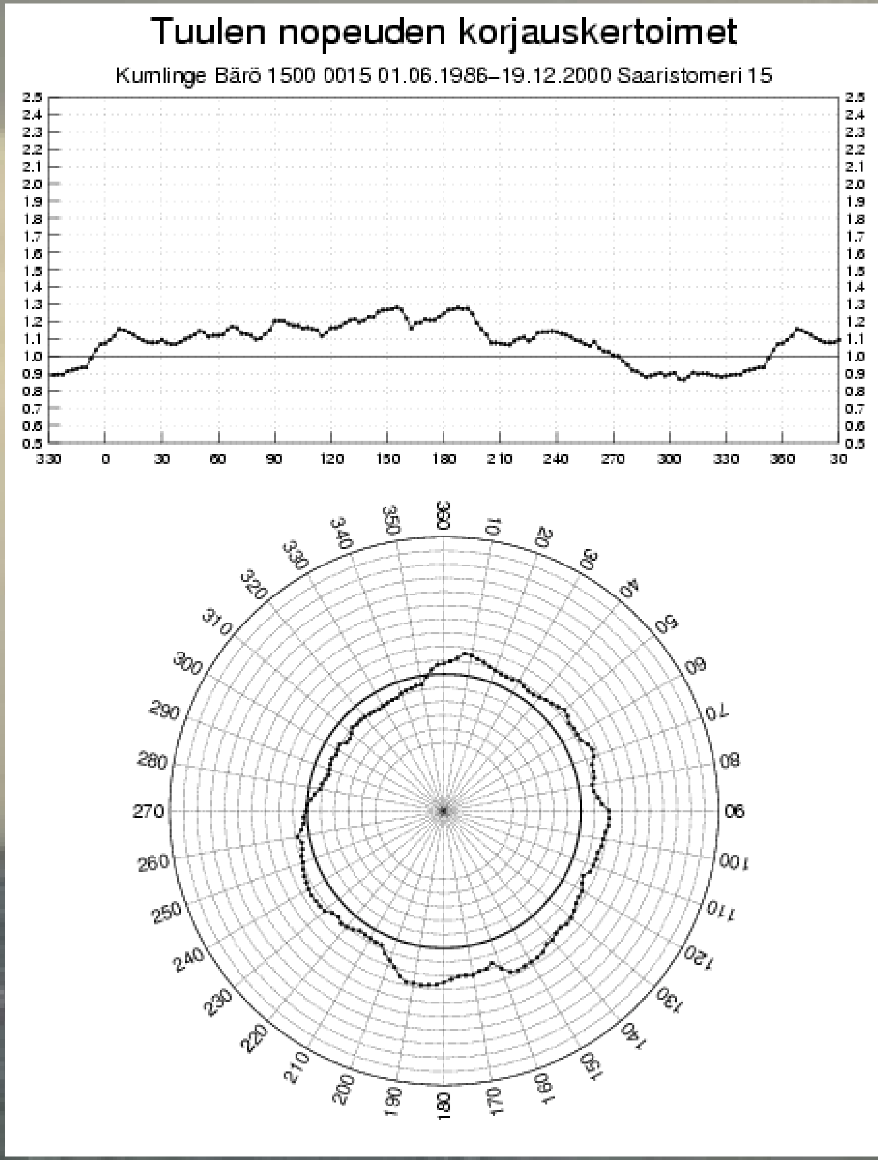
…even we have rare events & severe weather

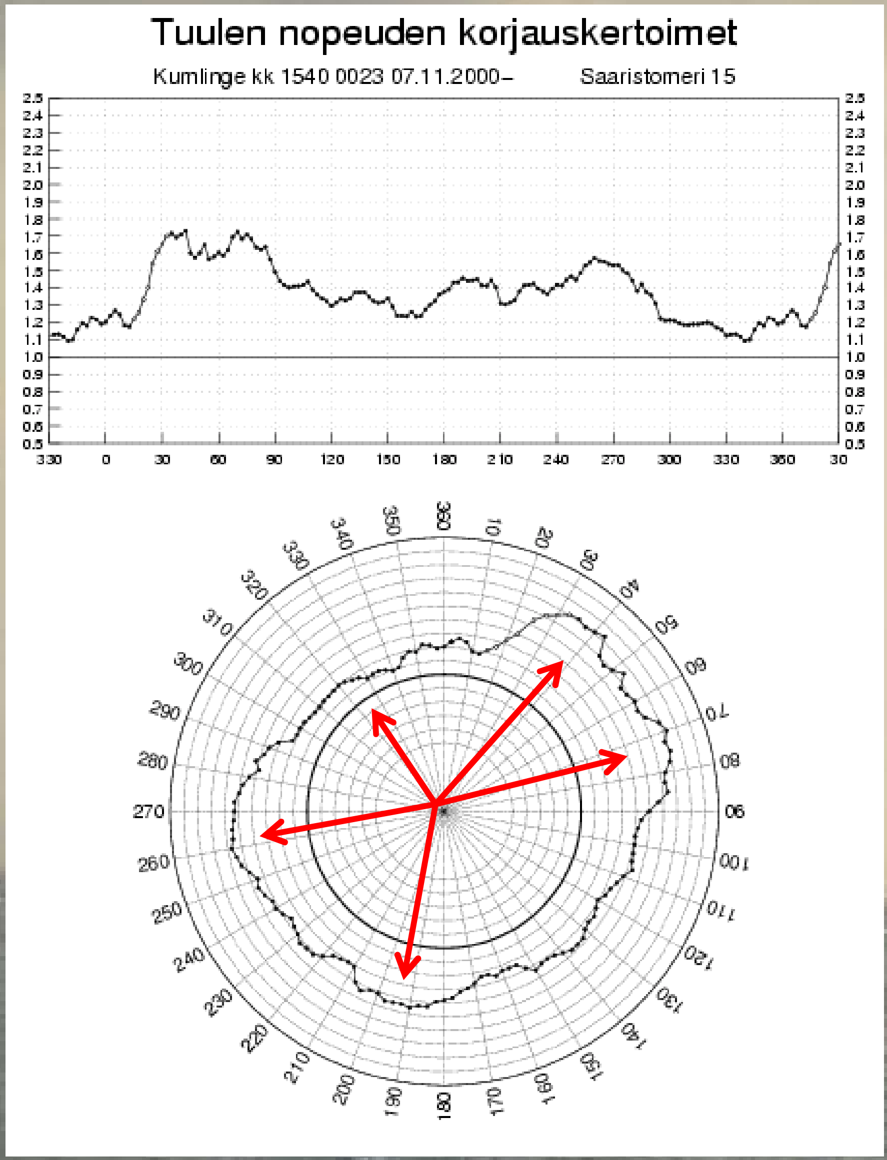




Case, FINLAND

…even we have rare events & severe weather





Case, FINLAND

Verification, Official targets

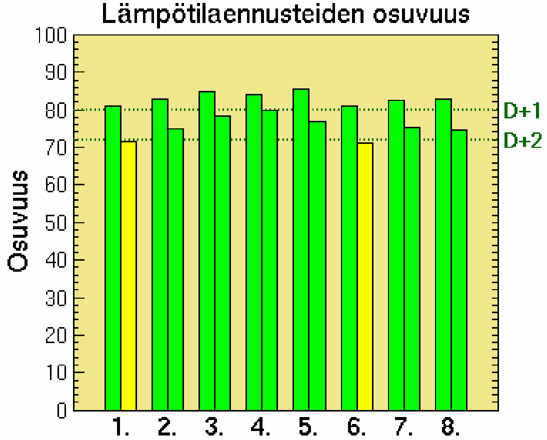
Viralliset verifiointitulokset

Lämpötila ▾ Osuvuus: Uusi (1999-) ▾ Jakso: 1.1 - 11.9 ▾ 2004 ▾

Laske | Info | Yhteenveto

Jakso: 1.1.2004 - 11.9.2004

		D+1	D+2
1.	KEPA –> Hki–Vantaa	81.2	71.5
2.	KEPA –> Tampere	82.8	74.9
3.	LSAP –> Tampere	84.7	78.3
4.	KEPA –> Sodankylä	84.0	79.8
5.	PSAP –> Sodankylä	85.5	77.0
6.	ISAP –> Kuopio	80.9	71.3
7.	Keskiarvo (1,2,4)	82.7	75.4
8.	Keskiarvo (1,3,5)	83.1	74.5



Tavoite vuodelle 1999: 80% yhdelle vuorokaudelle, 72% kahdelle. (merkitty katkoviivalla)

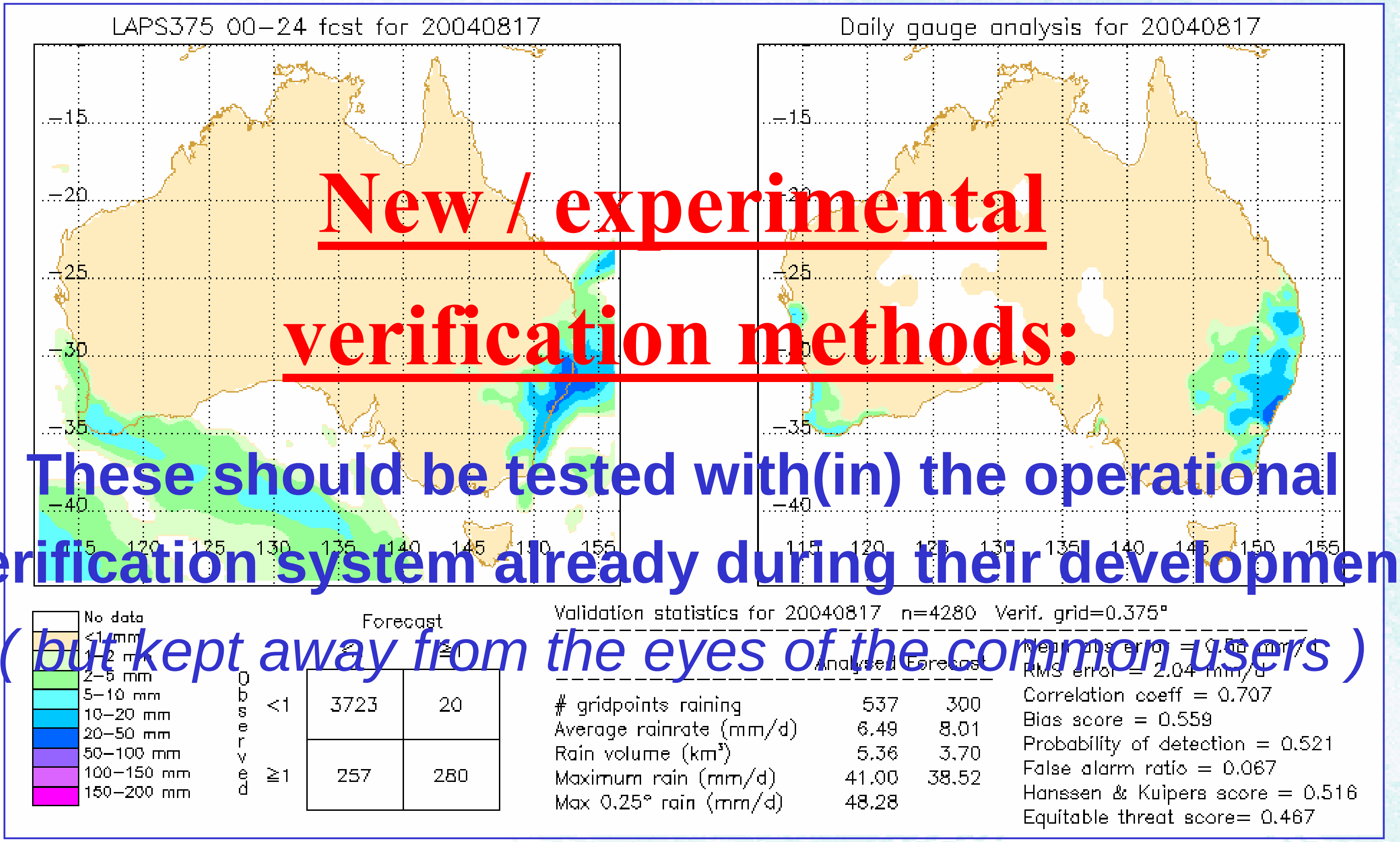
Uusi osuvuussuure mittaa alle annetun rajan olevien ennustusvirheiden osuutta.
Raja on 3 astetta Sodankylässä ja 2 astetta muualla Suomessa.

"hit calendar" → Osuuskalenteri: 1 2 3 4 5 6

time-series → Aikasarja: 1 2 3 4 5 6 7 8

Case, Australia *(Ebert et al.)*

Object-oriented verification: Daily rainfall - CRAs



New / experimental verification methods:

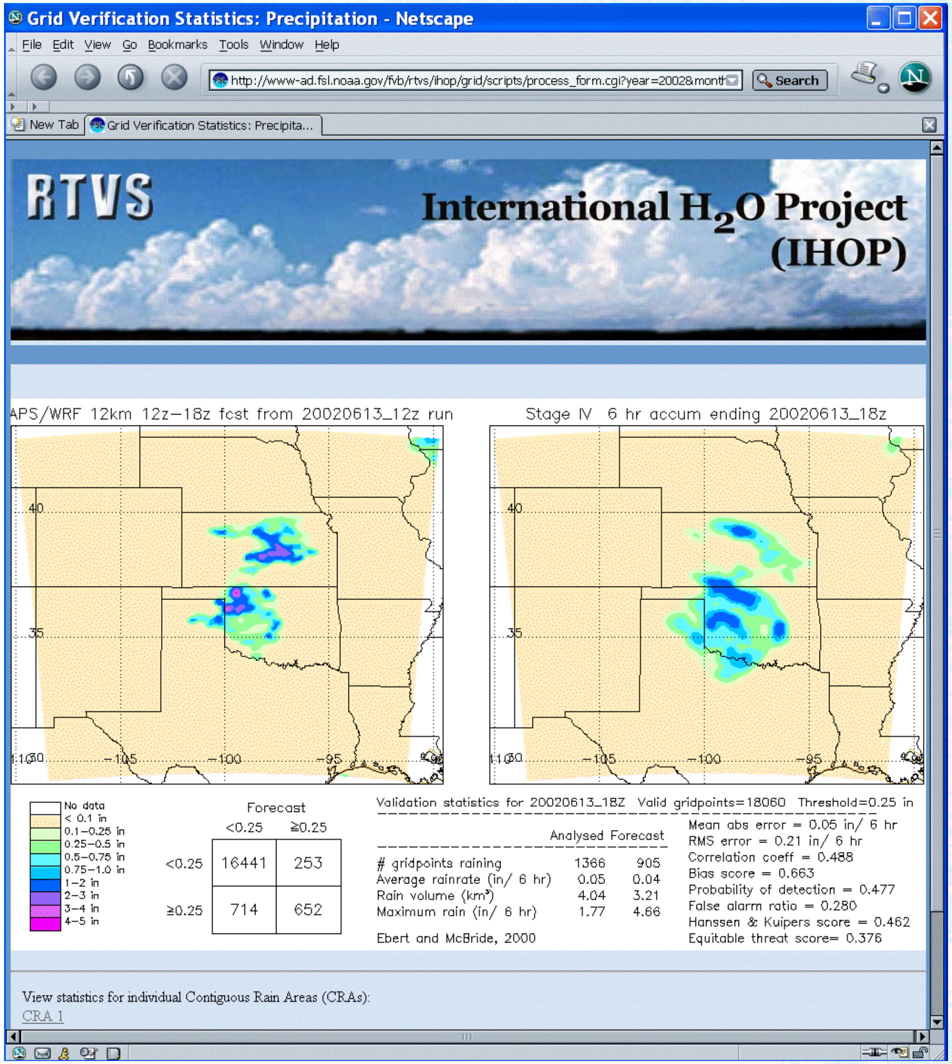
These should be tested with(in) the operational verification system already during their development

(but kept away from the eyes of the common users)

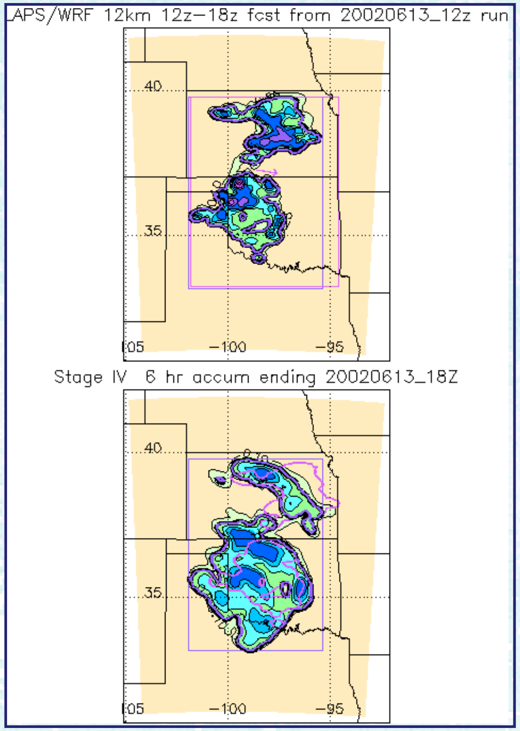
Verification of model QPF against gridded analysis of daily gauge data

Case, USA *(Ebert et al. & Loughe et al.)*

Object-oriented verification: Daily rainfall - CRAs



Implementation of *(Ebert and McBride)* technique to the WRF model evaluation



Specialized customer/product verification:

- ✓ Road maintenance authorities
- ✓ Energy production / consumption
- ✓ Construction
- ✓ Agriculture
- ✓ Hydrological applications
- ✓ Special occasional events, *e.g. Beijing Olympics*
- ✓ Golf courses
- ✓ Fire brigade
- ✓ etc...
- ✓ **All of these need special, customer-dependent, focus**
- ✓ **Feeback process both to the forecasters AND to the users, in real-time, linked to the service**
- ✓ **Aposteriori** (after service ended) **verification is of little value**
- ✓ **Methodology must be user-oriented !**

Case, Australia *(Ebert et al.)*

Fire weather elements

DAY+1 Errors on 20040330

Location	Fcs T	T(maxT)	Error	T(05Z)	Error	T(maxFDI)	Error	Fcs TD/RH	TD/RH(maxT)	Error	TD/RH(05Z)	Error	TD/RH(maxFDI)	Error
Armidale	25.0	24.4	+0.6	24.3	+0.7	24.4	+0.6	8.0	10.7	-2.7	10.6	-2.6	10.7	-2.7
Coffs Harbour	29.0	27.1	+1.9	25.3	+3.7	27.1	+1.9	16.0	15.8	+0.2	16.6	-0.6	15.8	+0.2
Grafton	29.0	-	-	-	-	-	-	14.0	-	-	-	-	-	-
Inverell	27.0	-	-	-	-	-	-	5.0	-	-	-	-	-	-
Jerrys Plains	33.0	-	-	-	-	-	-	10.0	-	-	-	-	-	-
Lismore	25.0	27.0	-2.0	27.0	-2.0	26.7	-1.7	16.0	14.8	+1.2	14.8	+1.2	14.5	+1.5

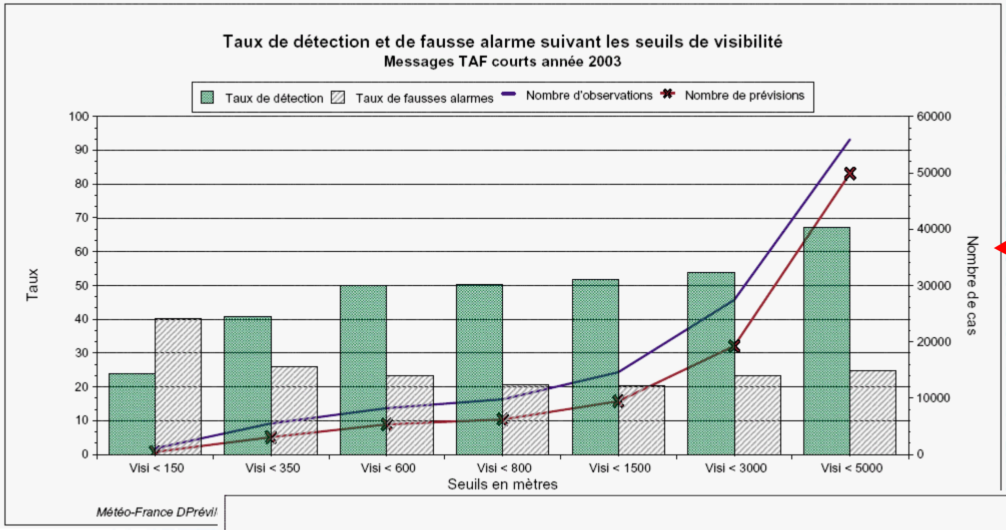
Location	Fcs WD	WD(maxT)	Error	WD(05Z)	Error	WD(maxFDI)	Error	Fcs WS	WS(maxT)	Error	WS(05Z)	Error	WS(maxFDI)	Error
Armidale	270.0	300.0	+1.0	290.0	+1.0	300.0	+1.0	15.0	20.5	-5.5	14.8	+0.2	20.5	-5.5
Coffs Harbour	23.0	10.0	+1.0	30.0	0.0	10.0	+1.0	25.0	27.7	-2.7	37.1	-12.1	27.7	-2.7
Grafton	45.0	-	-	-	-	-	-	20.0	-	-	-	-	-	-
Inverell	-	-	-	-	-	-	-	15.0	-	-	-	-	-	-
Jerrys Plains	293.0	-	-	-	-	-	-	25.0	-	-	-	-	-	-
Lismore	113.0	130.0	+1.0	130.0	+1.0	260.0	+7.0	30.0	7.6	+22.4	7.6	+22.4	9.4	+20.6

Location	Forrest Fire Danger Index							Grass of Moorland Fire Danger Index						
	Fcs FDI	FDI(maxT)	Error	FDI(05Z)	Error	FDI(maxFDI)	Error	Fcs FDI	FDI(maxT)	Error	FDI(05Z)	Error	FDI(maxFDI)	Error
Armidale	8.9	7.5	+1.4	6.5	+2.4	7.0	+1.9	-	-	-	-	-	-	-
Coffs Harbour	7.5	6.4	+1.1	5.5	+1.9	6.0	+1.5	0.0	0.0	-0.0	0.0	-0.0	0.0	+0.0
Grafton	8.0	-	-	-	-	-	-	-	-	-	-	-	-	-
Inverell	17.0	-	-	-	-	-	-	0.0	-	-	-	-	-	-
Jerrys Plains	23.4	-	-	-	-	-	-	0.0	-	-	-	-	-	-
Lismore	5.6	5.1	+0.6	5.1	+0.6	2.0	+3.6	-	-	-	-	-	-	-
Scone	26.3	13.0	+13.3	13.0	+13.3	13.0	+13.3	-	-	-	-	-	-	-
Taree	8.5	2.8	+5.7	3.5	+5.0	3.0	+5.5	0.0	0.0	+0.0	0.0	+0.0	0.0	+0.0

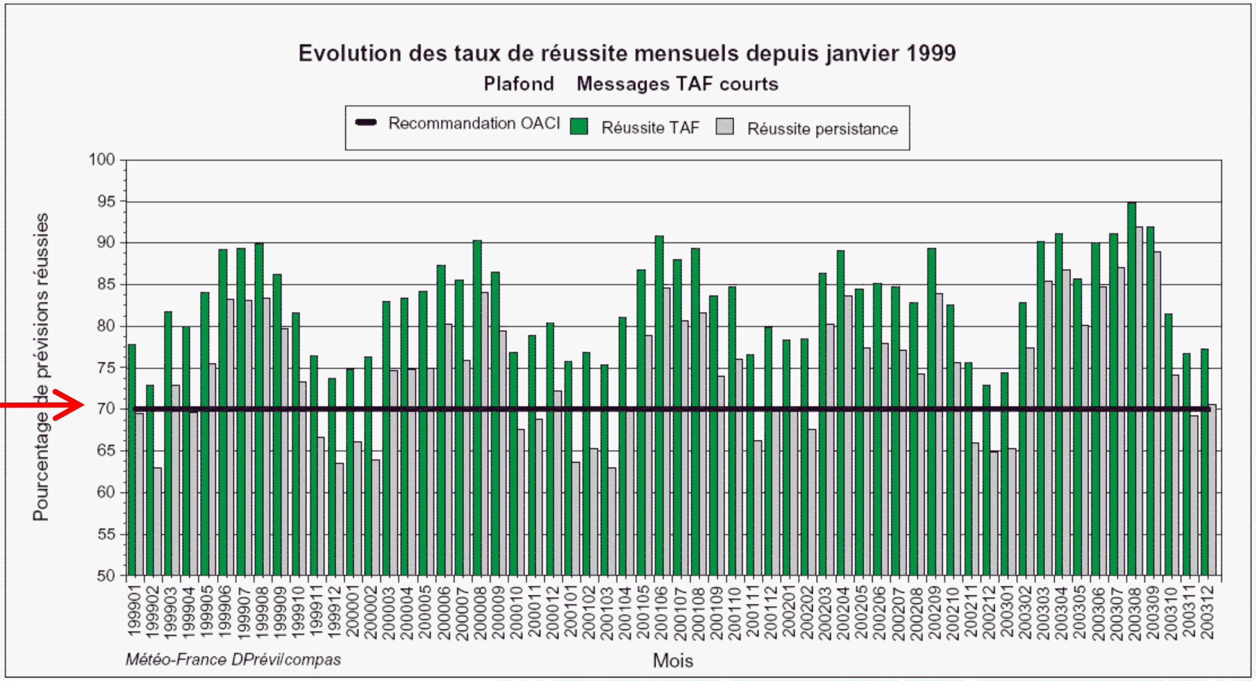
Specialized verification: Aviation, TAF

- ✓ “Totally it‘s own world“
- ✓ Often separate from „terrestrial“ forecast production... Possibly requiring an independent verification system
- ✓ Similar prerequisites as for all operational verification
- ✓ Not elaborated here, except with one example

Case, FRANCE *(Atger)*



POD / FAR ; Depend on p(E)



User (ICAO) -defined target score

Techniques of verification systems

- ✓ **Personally, I'm a novice of modern IT techniques !**

  => No answers on "How to do it?" in this talk
- ✓ **Well-structured, rapid, interactive user interface and verification database**
- ✓ **How about, e.g. packages like "R" ???**
  - Are they easy to link to various, different local database structures?
- ✓ **Comprehensive user-guide/tutorial "tool bag"**

Verification "toolbag"

Continuous variables, *Exploring the data*

- ✓ **Scatterplots of forecasts vs. observations**
  - – Visual relation between forecast and observed distributions
  - – Distinguish outliers in forecast and/or observation datasets
- ✓ **Additional scatterplots**
  - – Observations vs. [forecast - observation] *difference*
  - – Forecasts vs. [forecast - observation] *difference*
  - – Behaviour of forecast errors with respect to observed or forecast distributions - potential clustering or curvature in the relationships
- ✓ **Time-series plot of forecasts vs. observations (or forecast error)**
  - – Potential outliers in either forecast or observation datasets
  - – Trends and time-dependent relationships

Verification tutorial, user-guide "toolbag" of an operational system

might include something like what follows,

time permitting, probably not.

Verification "toolbag"

Continuous variables:

Mean Error aka **Bias** : $ME = (1/n) \sum (f_i - o_i)$

Mean Absolute Error: $MAE = (1/n) \sum | f_i - o_i |$

(Root) Mean Squared Error: $(R)MSE = (1/n) \sum (f_i - o_i)^2$

(General) Skill Score: $SS = (A - A_{ref}) / (A_{perf} - A_{ref})$

$SS = [1 - A / A_{ref}] * 100$

$MAE\_SS = [1 - MAE / MAE_{ref}] * 100$

$MSE\_SS = [1 - MSE / MSE_{ref}] * 100$

Latter also known as **Reduction of Variance, RV**

Linear Error in Probability Space:

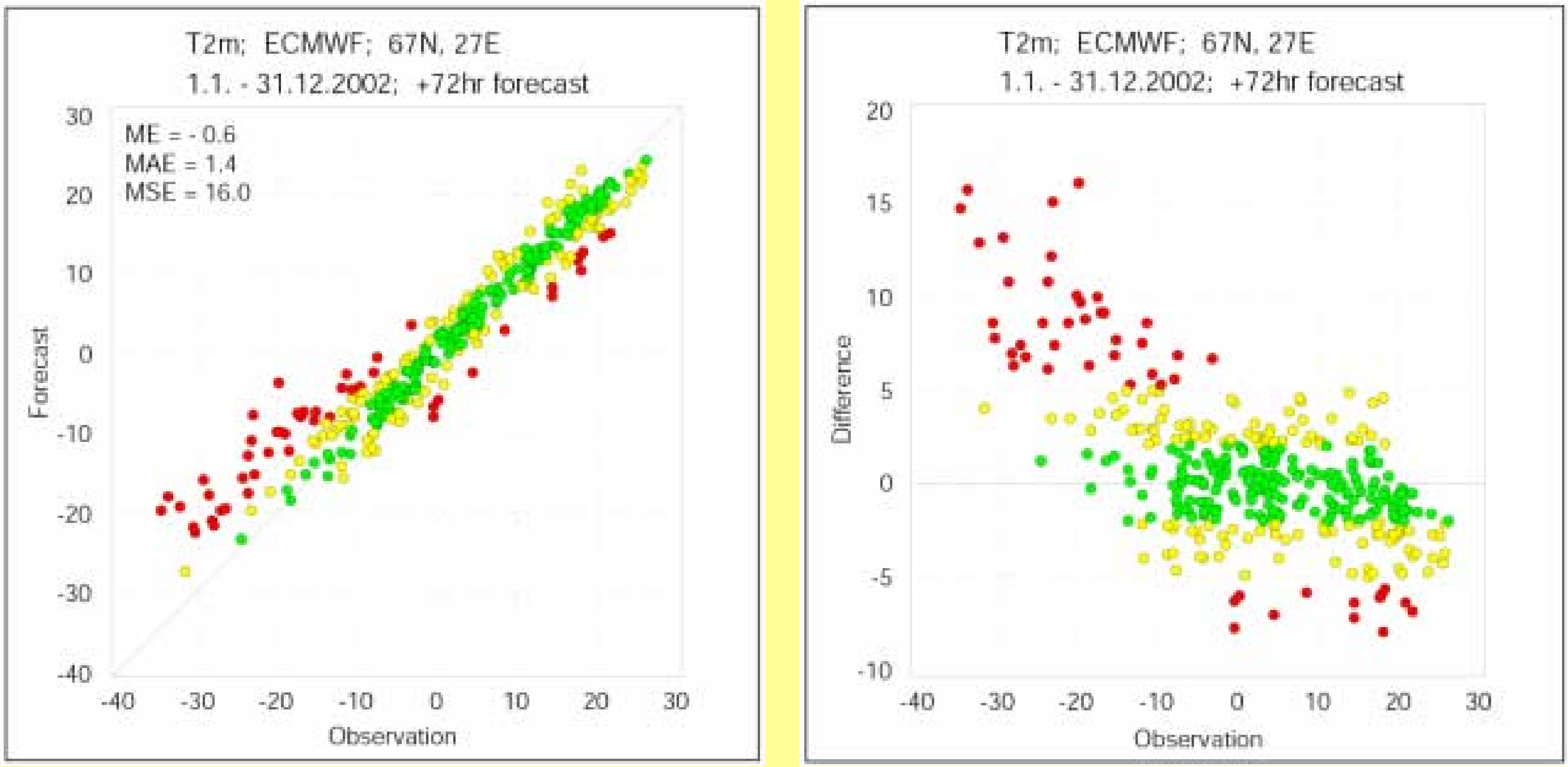
$LEPS = (1/n) \sum | CDF_o (f_i) - CDF_o (o_i) |$

LEPS Skill Score:

$LEPS\_SS = [1 - LEPS / LEPS_{ref}] * 100$

Verification "toolbag"

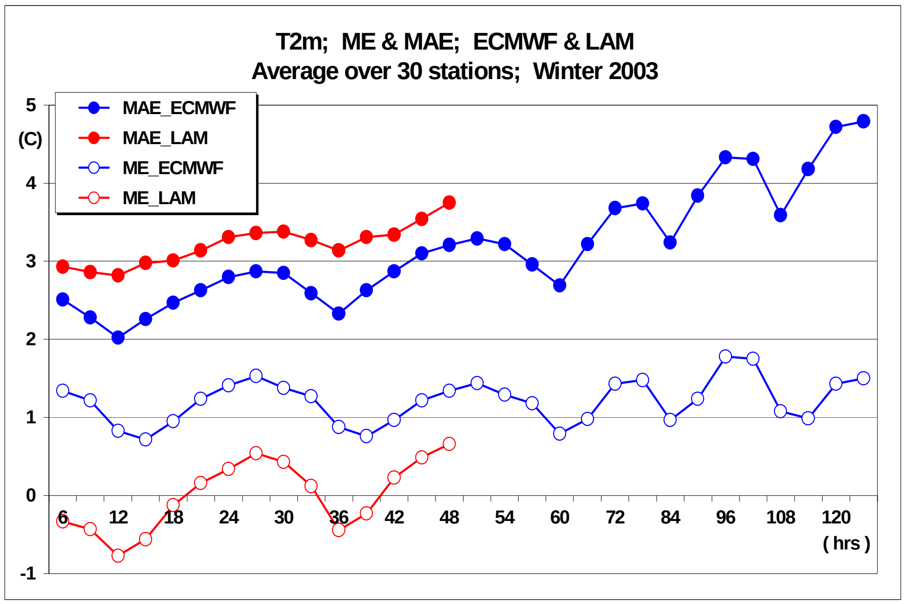
Continuous Variables - Example 1; Exploring the data

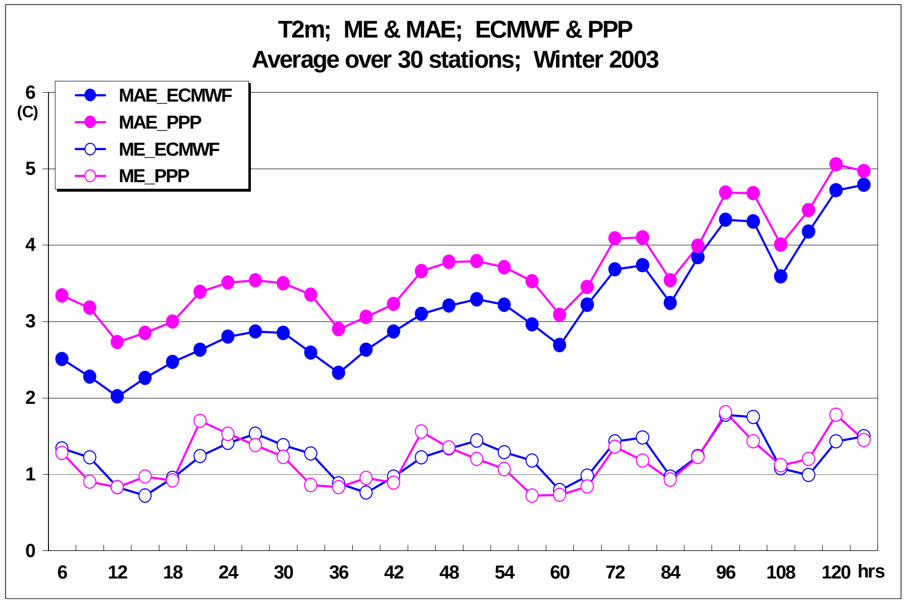


Scatterplot of one year of ECMWF three-day T2m forecasts (left) and forecast errors (right) versus observations at a single location. Red, yellow and green dots separate the errors in three categories. Some basic statistics like ME, MAE and MSE are also shown. The plots reveal the dependence of model behaviour with respect to temperature range, i.e. over- (under-) forecasting in the cold (warm) tails of the distribution.

Verification "toolbag"

Continuous Variables - Example 2

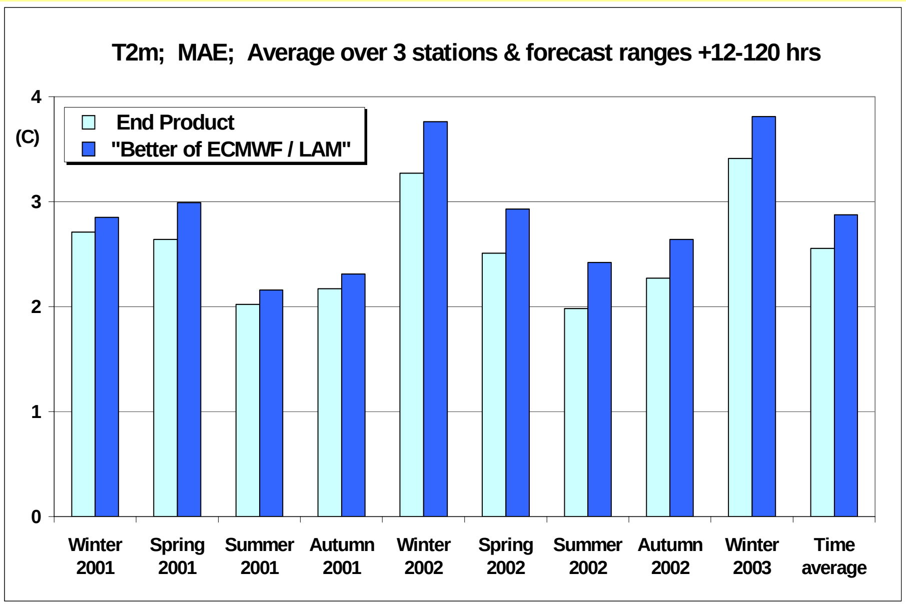


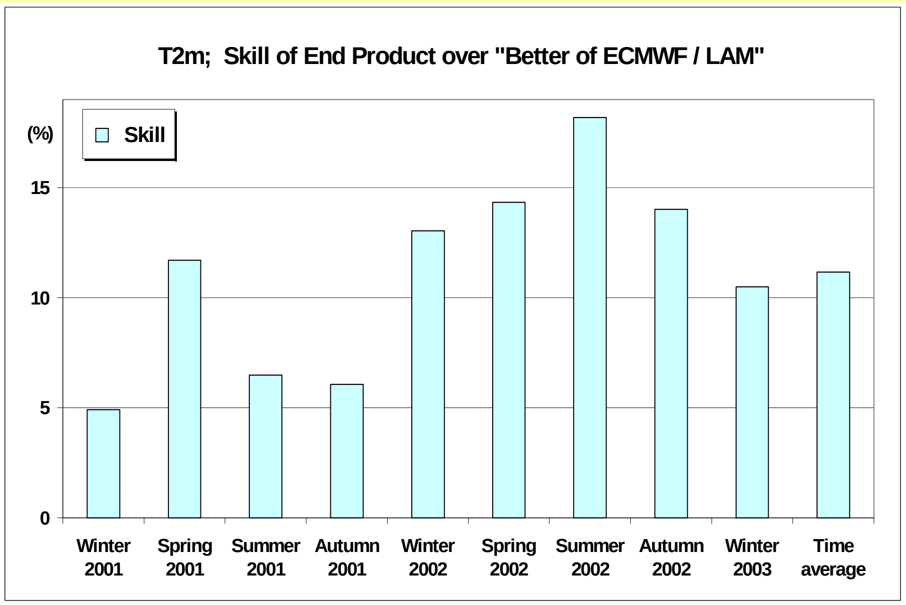


Temperature bias and MAE comparison between ECMWF and a Limited Area Model (LAM) (left), and an experimental post-processing scheme (PPP) (right), aggregated over 30 stations and one winter season. In spite of the ECMWF warm bias and diurnal cycle, it has a slightly lower MAE level than the LAM (left). The applied experimental "perfect prog" scheme does not manage to dispose of the model bias and exhibits larger absolute errors than the originating model – this example clearly demonstrates the importance of thorough verification prior to implementing a potential post-processing scheme into operational use.

Verification "toolbag"

Continuous Variables - Example 3

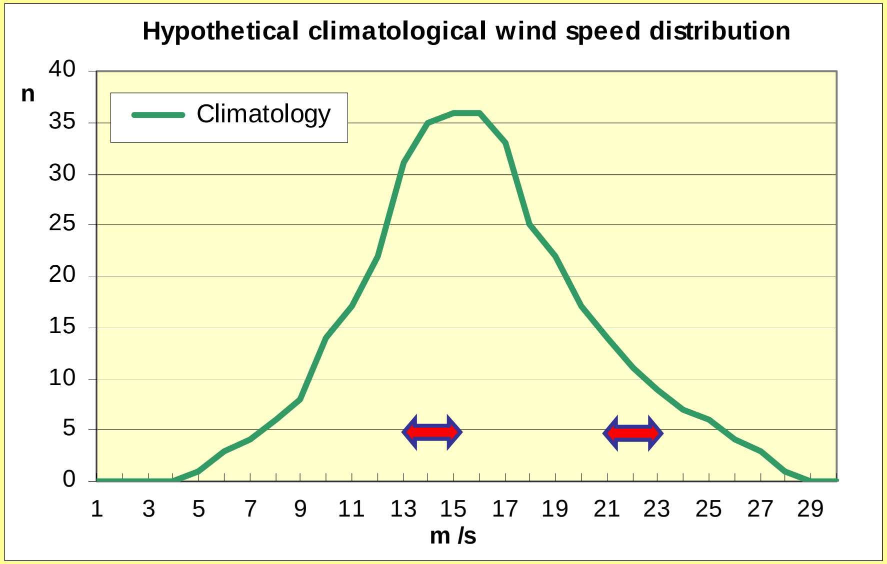


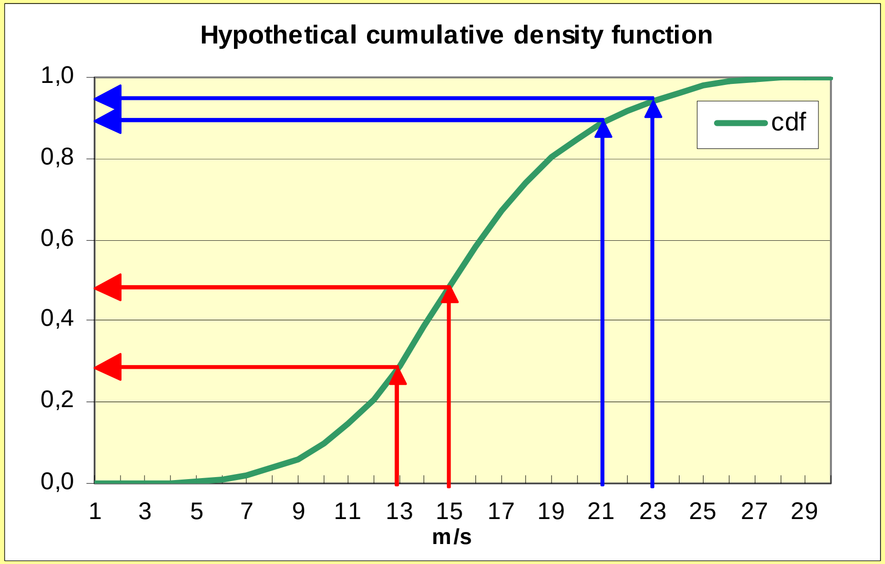


Mean Absolute Errors of End Product and DMO temperature forecasts (left), and Skill of the End Products over model output (right). The better of either ECMWF or local LAM is chosen up to the +48 hour forecast range (hindcast), thereafter ECMWF is used. The figure is an example of both aggregation (3 stations, several forecast ranges, two models, time-average) and stratification (seasons).

Verification "toolbag"

Continuous Variables - Example 4

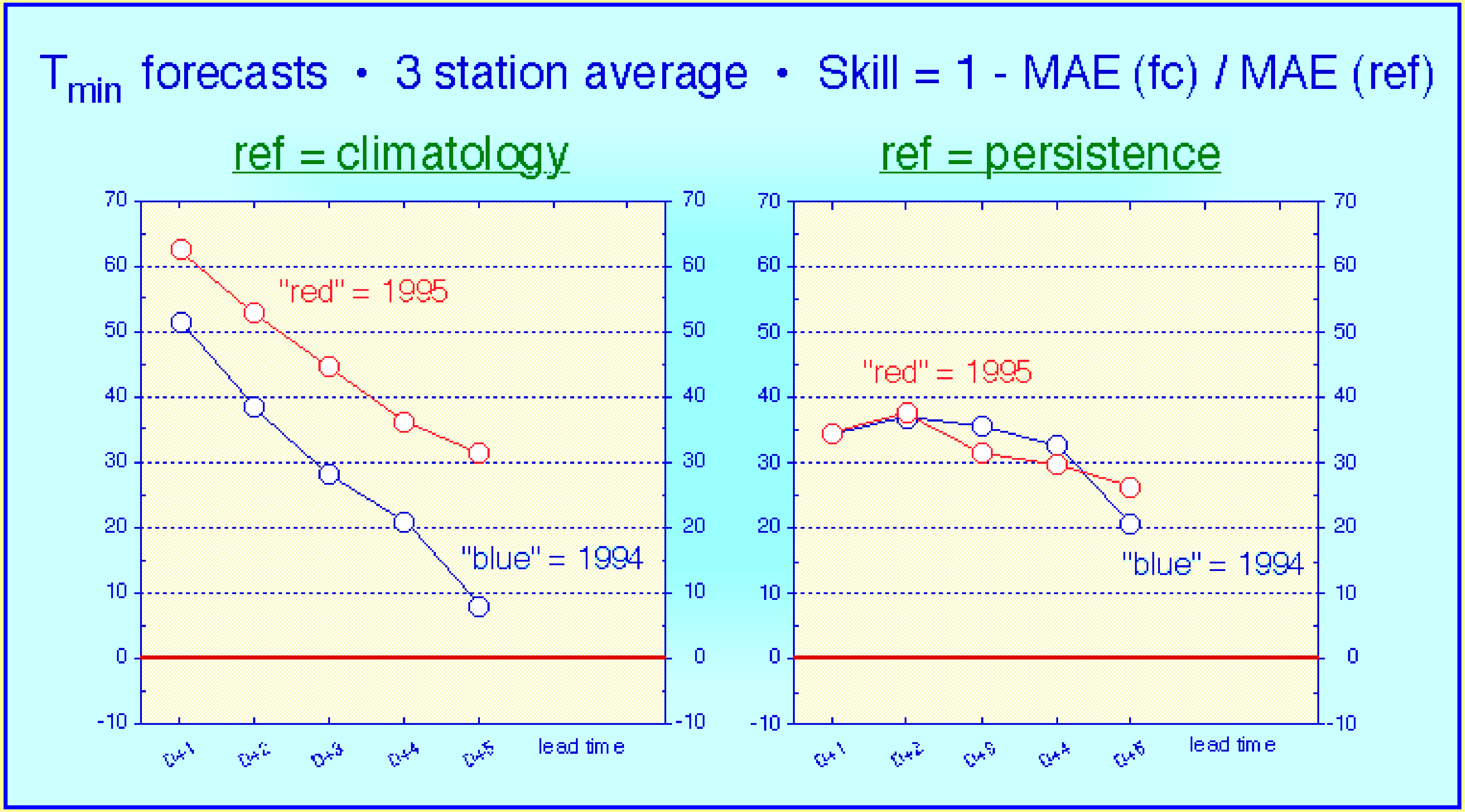




Application and computation of LEPS for a hypothetical wind speed distribution at an assumed location, where the climatological frequency distribution (left) is transformed to a cumulative probability distribution (right). A 2 m/s forecast error around the median, in the example 15 m/s vs. 13 m/s (red arrows), would yield a LEPS value of c. 0.2 in the probability space (| 0.5 – 0.3 |, red arrows). However, an equal error in the measurement space close to the tail of the distribution, 23 m/s vs. 21 m/s (blue arrows), would result a LEPS value of c. 0.05 (| 0.95 – 0.9 |, blue arrows). Hence forecast errors of rare events are much less penalized using LEPS.

Verification "toolbag"

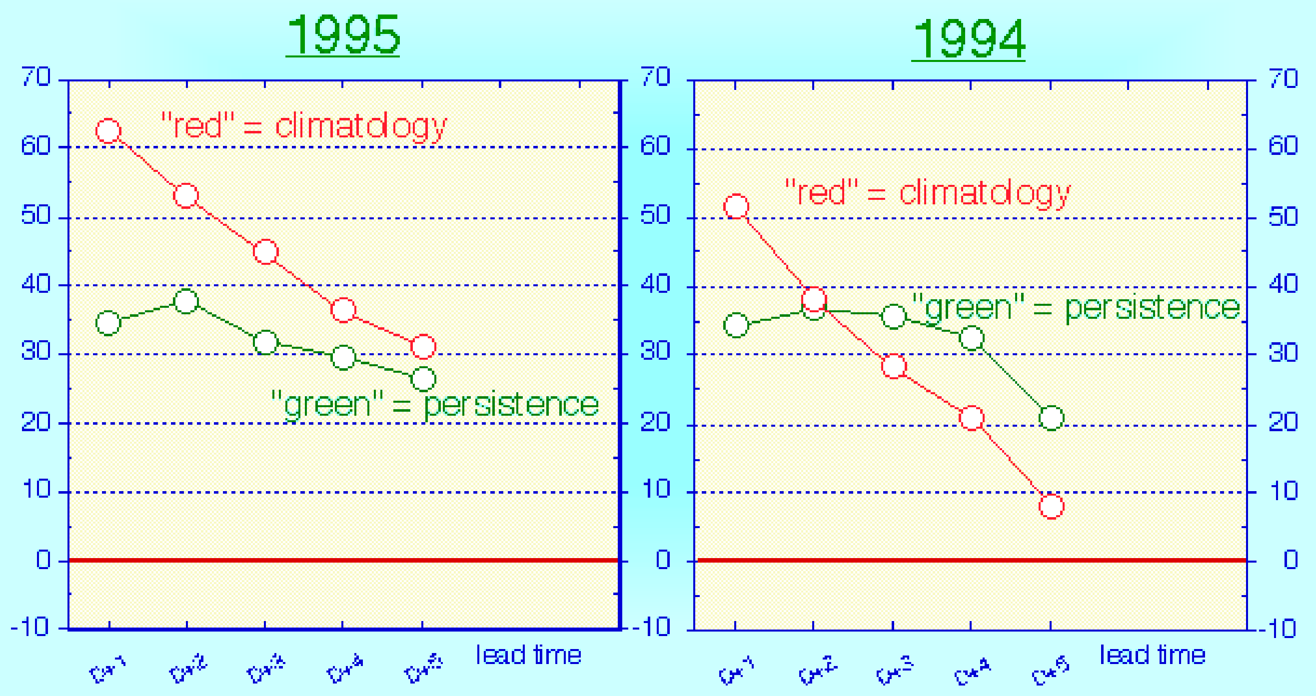
Skill comparison (example $_1$) ...



Verification "toolbag"

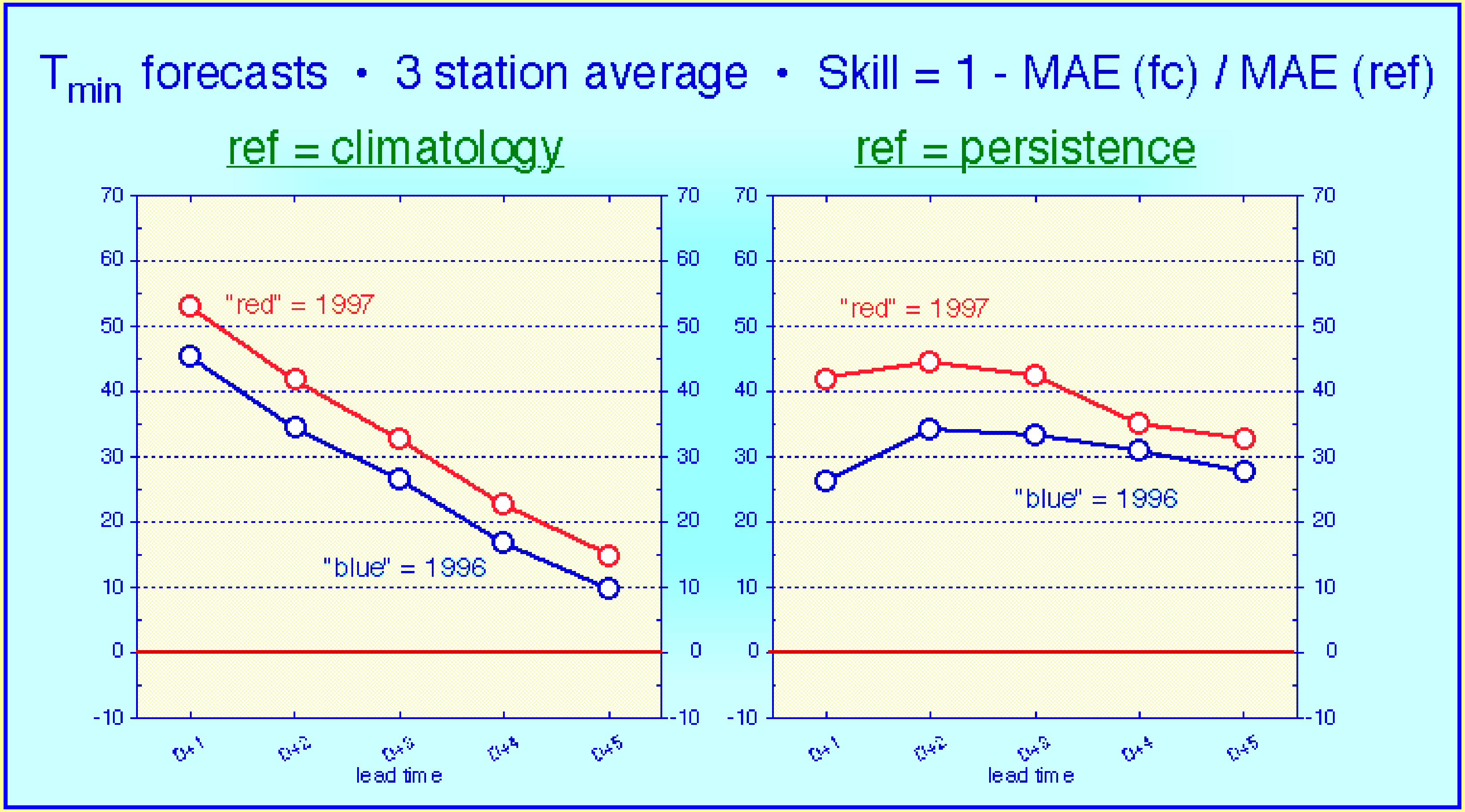
Skill comparison (example $_2$) ...

T_{min} forecasts • 3 station average • Skill = 1 - MAE (fc) / MAE (ref)



Verification "toolbag"

Skill comparison (example $_3$) ...



Verification "toolbag"

Categorical Events

Event forecast	Event observed		
	Yes	No	Marginal total
Yes	Hit	False alarm	Fc Yes
No	Miss	Corr. rejection	Fc No
Marginal total	Obs Yes	Obs No	Sum total

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

Verification "toolbag"

Categorical Events

Bias = Frequency Bias Index B = FBI = (a + b) / (a + c)

Proportion Correct PC = (a + d) / n

Probability Of Detection, Hit Rate (H), Prefigurance

POD = a / (a + c)

False Alarm Ratio FAR = b / (a + b)

Post agreement PAG = a / (a + b)

False Alarm Rate, Probability of False Detection (POFD)

F = b / (b + d)

Hanssen & Kuiper's Skill Score, True Skill Statistics

KSS = TSS = POD – F

= (ad – bc) / [(a+c) (b+d)]

Threat Score, Critical Success Index TS = CSI = a / (a + b + c)

Equitable Threat Score ETS = $(a - a_r) / (a + b + c - a_r)$

Heidke Skill Score HSS = 2 (ad–bc) / [(a+c)(c+d) + (a+b)(b+d)]

Odds ratio OR = a*d / b*c

ORSS = (ad–bc) / (ad+bc) = (OR–1) / (OR+1)

Verification "toolbag"

Rain forecast	Rain observed		
	Yes	No	fc Σ
Yes	52	45	97
No	22	227	249
obs Σ	74	272	346

~>

Precipitation example

B	= 1.31	TS	= 0.44
PC	= 0.81	ETS	= 0.32
POD	= 0.70	KSS	= 0.53
FAR	= 0.46	HSS	= 0.48
PAG	= 0.54	OR	= 11.92
F	= 0.17	ORSS	= 0.85

Generalization of KSS and HSS – measures of improvement over random forecasts:

$$KSS = \{ \Sigma p (f_i , o_i) - \Sigma p (f_i) p (o_i) \} \ / \ \{ 1 - \Sigma (p (f_i))^2 \}$$

$$HSS = \{ \Sigma p (f_i , o_i) - \Sigma p (f_i) p (o_i) \} \ / \ \{ 1 - \Sigma p (f_i) p (o_i) \}$$

Clouds forecast	Clouds observed			
	0 - 2	3 - 5	6 - 8	fc Σ
0 - 2	65	10	21	96
3 - 5	29	17	48	94
6 - 8	18	10	128	156
obs Σ	112	37	197	346

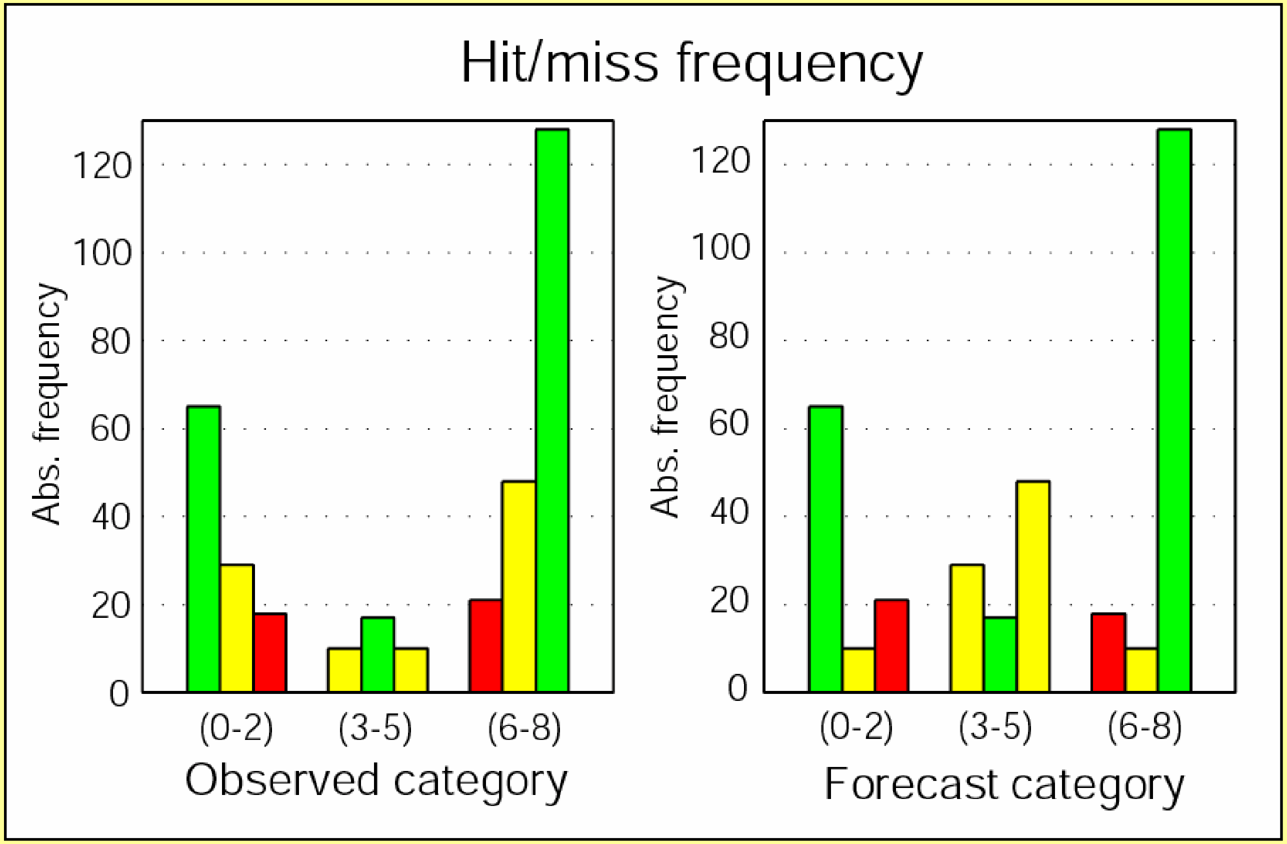
~>

No clouds (0-2)	Partly cloudy (3-5)	Cloudy (6-8)
B = 0.86	B = 2.54	B = 0.79
POD = 0.58	POD = 0.46	POD = 0.65
FAR = 0.32	FAR = 0.82	FAR = 0.18
F = 0.13	F = 0.25	F = 0.19
TS = 0.45	TS = 0.15	TS = 0.57

Overall: PC = 0.61 KSS = 0.41 HSS = 0.37

Verification "toolbag"

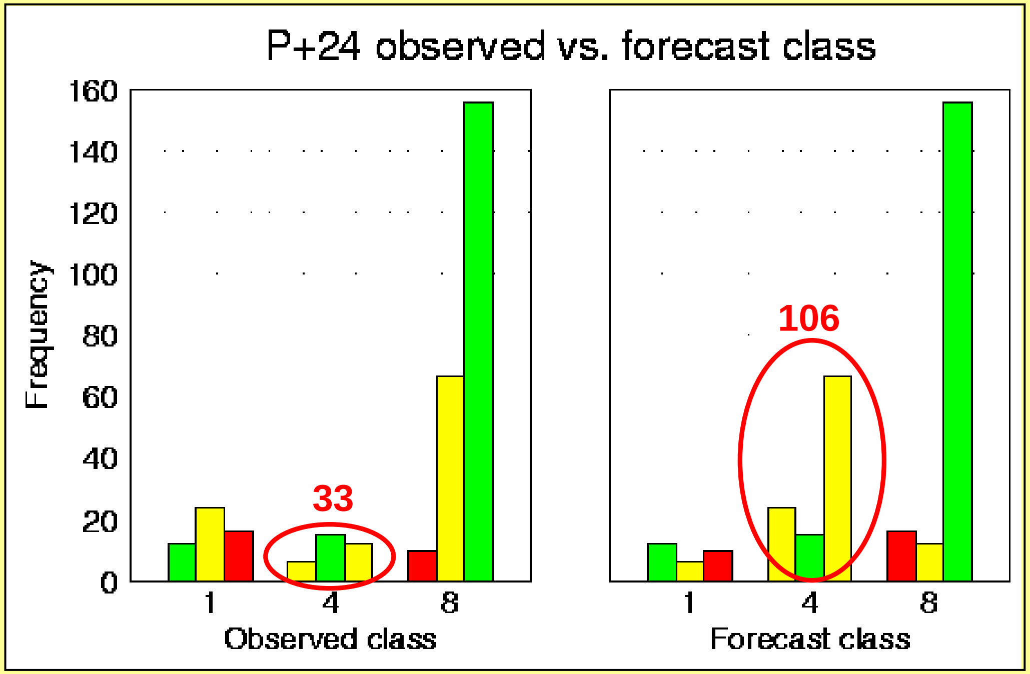
Categorical Events



The previous data transformed into hit/miss bar charts, either given the observations (left), or given the forecasts (right). The green, yellow and red bars denote correct and one and two category errors, respectively. The U-shape in observations is clearly visible (left), whereas there is no hint of such in the forecast distribution (right).

Verification ”toolbag”

Categorical Events



Observation

Forecaster	0–2	3–5	6–9	
1	12	6	10	28
4	24	15	67	106
8	16	12	156	184
	52	33	233	**318**

	Class		
	1	4	8
BIAS	0.5	3.2	0.8
TS	17.6	12.1	59.8
POD	23.1	45.5	67.0
FAR	57.1	85.8	15.2

PC = 57.5
KSS = 24.1

Verification ”toolbag”

Probability Forecasts

- ✓ All forecasting involves some level of uncertainty
- ✓ Deterministic forecasts cannot address the inherent uncertainty of the weather parameter or event
- ✓ Conversion of probability forecasts to categorical events is simple by defining the “on/off” threshold. Reverse is not straightforward.
- ✓ Verification is somewhat laborious => Large datasets are required to obtain any significant information

Verification "toolbag"

Probability Forecasts: Measures

Brier Score $BS = (1/n) \; \Sigma \; (p_i - o_i)^2$

Brier Skill Score $BSS = [1 - BS / BS_{ref}] * 100$

Ranked Probability Score $RPS = (1/(k-1)) \; \Sigma \; \{ (\Sigma p_i) - (\Sigma o_i) \}^2$

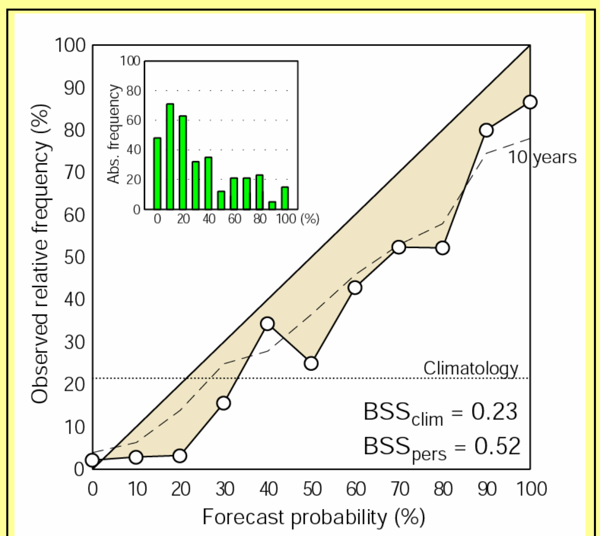
Ranked Probability Skill Score $RPSS = [1 - RPS / RPS_{ref}] * 100$

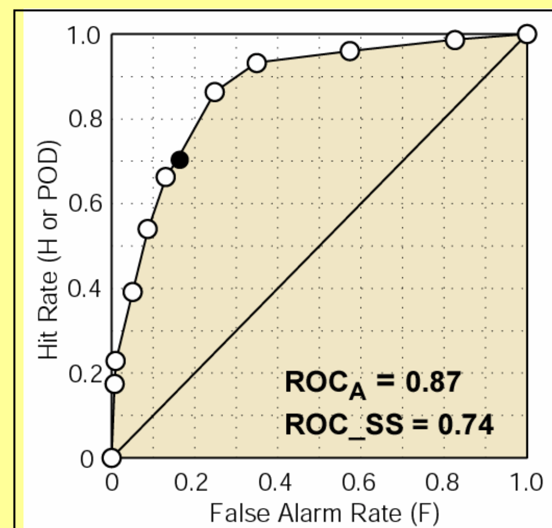
ROC_A Area based skill score: $ROC\_SS = 2 * ROC_A - 1$

ROC Curve

Reliability or

Attributes Diagram





Operational verification systems in Europe 2003:

My best guess…

(Based on ECMWF "Green book" Member State verification reporting)

	End Prod verif.	EPS verif.	EFI verif.	Subj. verif.	Season verif.	Wave height verif.	Other / Special:
1 Austria							Bi-annual verification reports intranet (?)
2 Belgium							
3 Croatia	+				(2*2)	+	
4 Czech							
5 Denmark	+	Talagrand			(3*3)		
6 Finland	+						On-line verif. package intranet; Periodical verifcation reports
7 France	+	+		4 class		+	
8 Germany	+	+					
9 (Greece)							
10 Hungary	+			3 class	(3*3)		On-line verif. package intra/internet
11 Iceland							On-line verif. package intranet;
12 Ireland							
13 Italy							
14 (Luxembourg)							
15 Netherlands	+						
16 Norway				+			On-line verification package Periodical verification reports;
17 Portugal							
18 (Serbia & MN)							
19 Slovenia							
20 Spain	?	+		+	+		Monthly reports
21 Sweden				5 class			Periodical verification reports;
22 Switzerland		+	+				
23 Turkey							
24 U. K.	?	+		8 class	+	+	On-line verif. package intranet - not reported;
Total	7 / 9	6	1	6	5	3	

Conclusions re. Operational Verification System(s)

- ✓ **Must cover the user AND the scientific perspective**
- ✓ **Customers/users:**
  - Forecasters
  - R&D modelers & developers
  - Various customers, users, the general public
  - Administration, decision-makers
- ✓ **Consistent package, must be made to last**
- ✓ **Not for "background backtalk", but in the FOREFRONT of everyday operational forecasting practice**
- ✓ **Tempting, user-friendly interface –> Rewarding to use**
- ✓ **Kept "operationally" up-to-data, e.g. no "loose links"**

Thank You
for bearing with me !!!