

Probability and ensemble forecasts: how do we evaluate their skill?

(an operational centre perspective)

Anna Ghelli, ECMWF

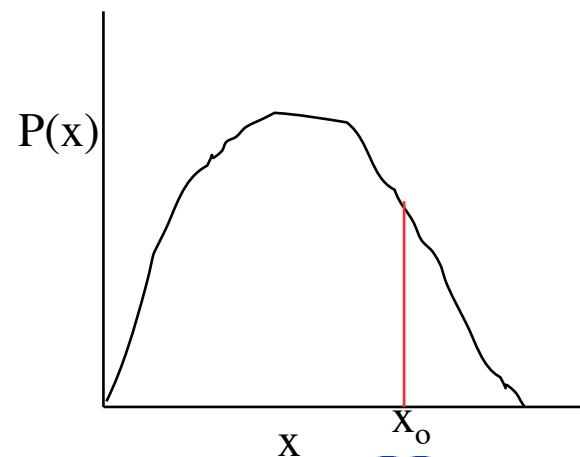
Thanks to Martin Janousek

Outline

- **Strategy for ensemble verifications**
- **Metrics**
- **The issues: who are the users?**
- **Examples of verification plots**
- **Flow dependent verification**
- **Introducing observation uncertainty**
- **Conclusions**
- **My questions**

Basics

- Verification of the ensemble pdf:
 - CRPS --> RPSS
 - Talagrand diagram
- Verification for pdfs of generic probability forecasts:
 - Ignorance Score
- Verification of forecasts of the probability of an event:
 - Brier Skill Score
 - ROC area
 - Reliability diagrams



Scores -- what they measure

- **Measuring bias**
 - Reliability diagram
- **Measuring total error**
 - Brier score (analogous to MSE) --> Brier Skill Score
- **Measuring potential skill**
 - Relative operating characteristic (ROC) --> ROC area
- **Measuring accuracy**
 - Ranked probability score --> RPSS
- **Measuring value**
 - Ignorance score $IGN = -\log_2 f_i$ --> ignorance skill score

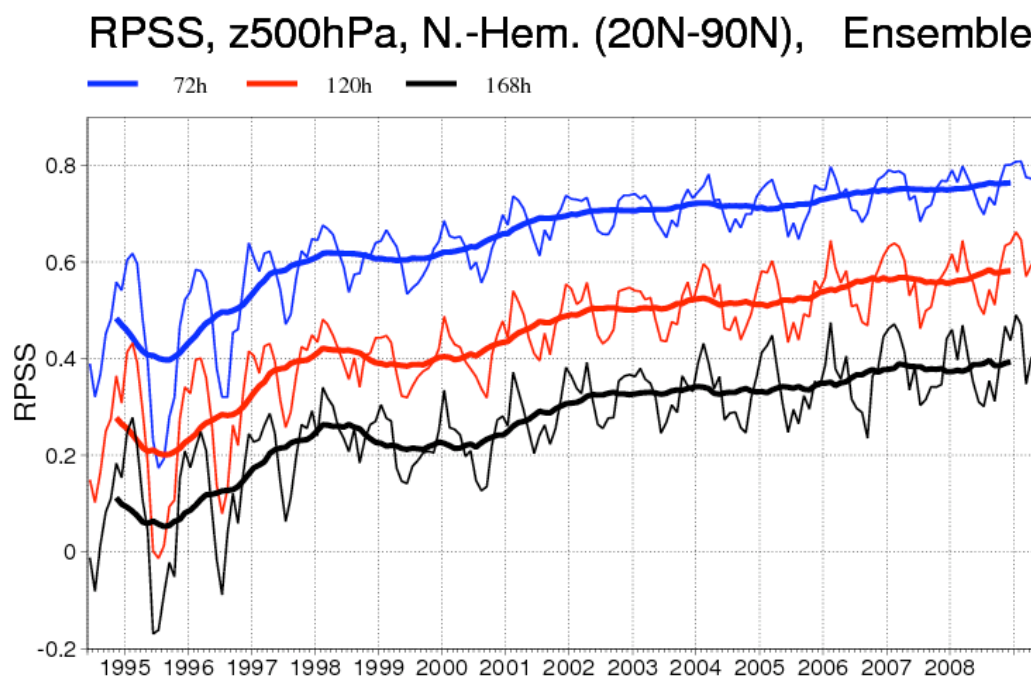
Verification questions

- What do the users want to know?
 - ◆ Administrators: Are we meeting the targets set in the annual plan?
 - ◆ Modelers: are the latest changes in the model algorithm improving the system?
 - ◆ Forecasters: Do I trust the forecast probabilities at their face value?
- How should verification results be used to improve daily forecasts?
- How should we look at these verification results?
 - ◆ Long term verification
 - ◆ Case studies

Looking at the long term improvements

The administrator's perspective:
Are the set targets going to be reached?

Target: xx days in predictability



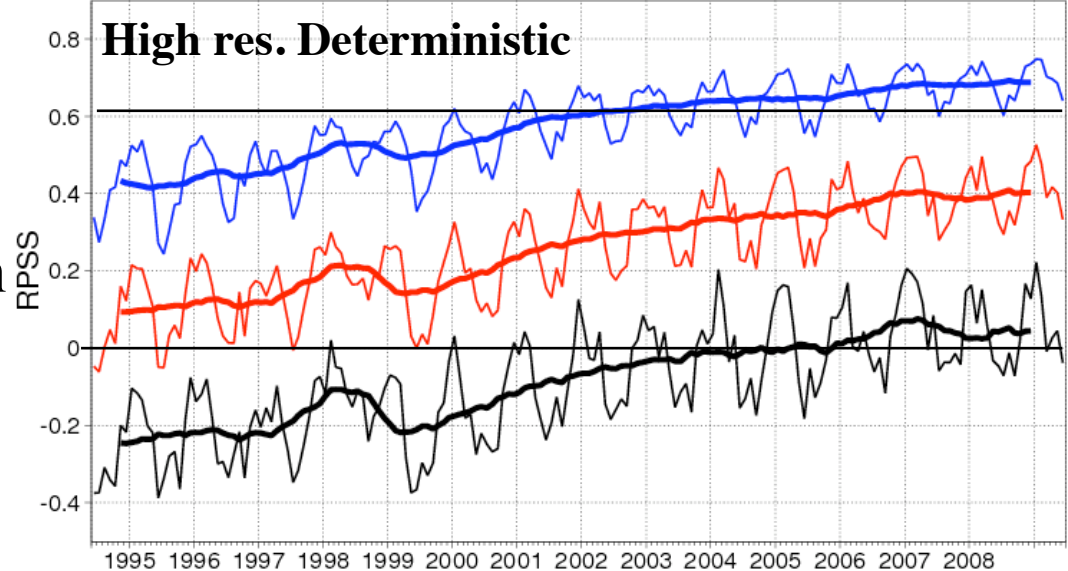
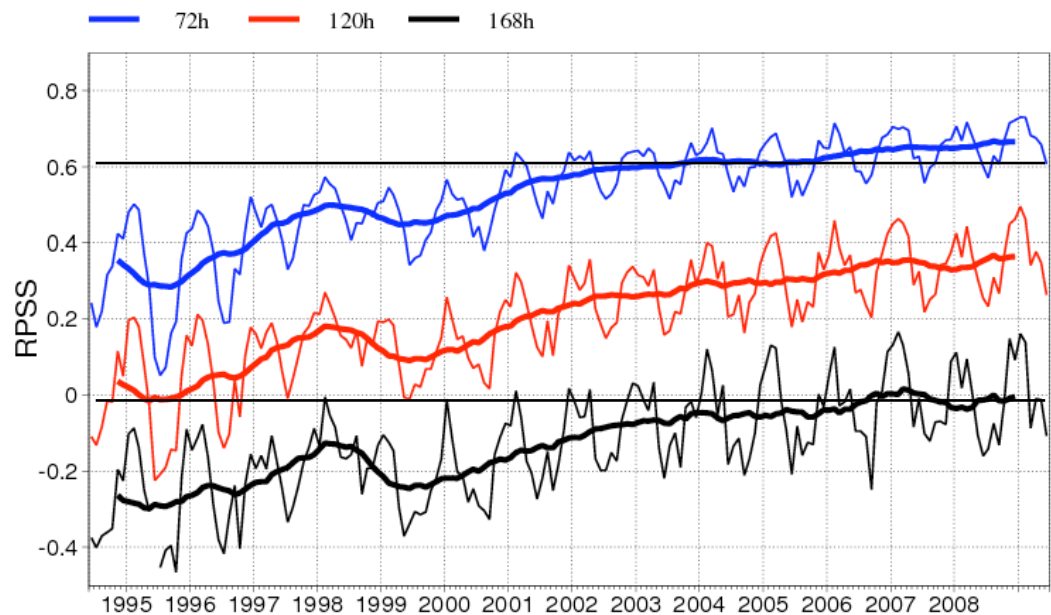
Looking at the long term improvements

The administrator's perspective:

Are the set targets going to be reached?

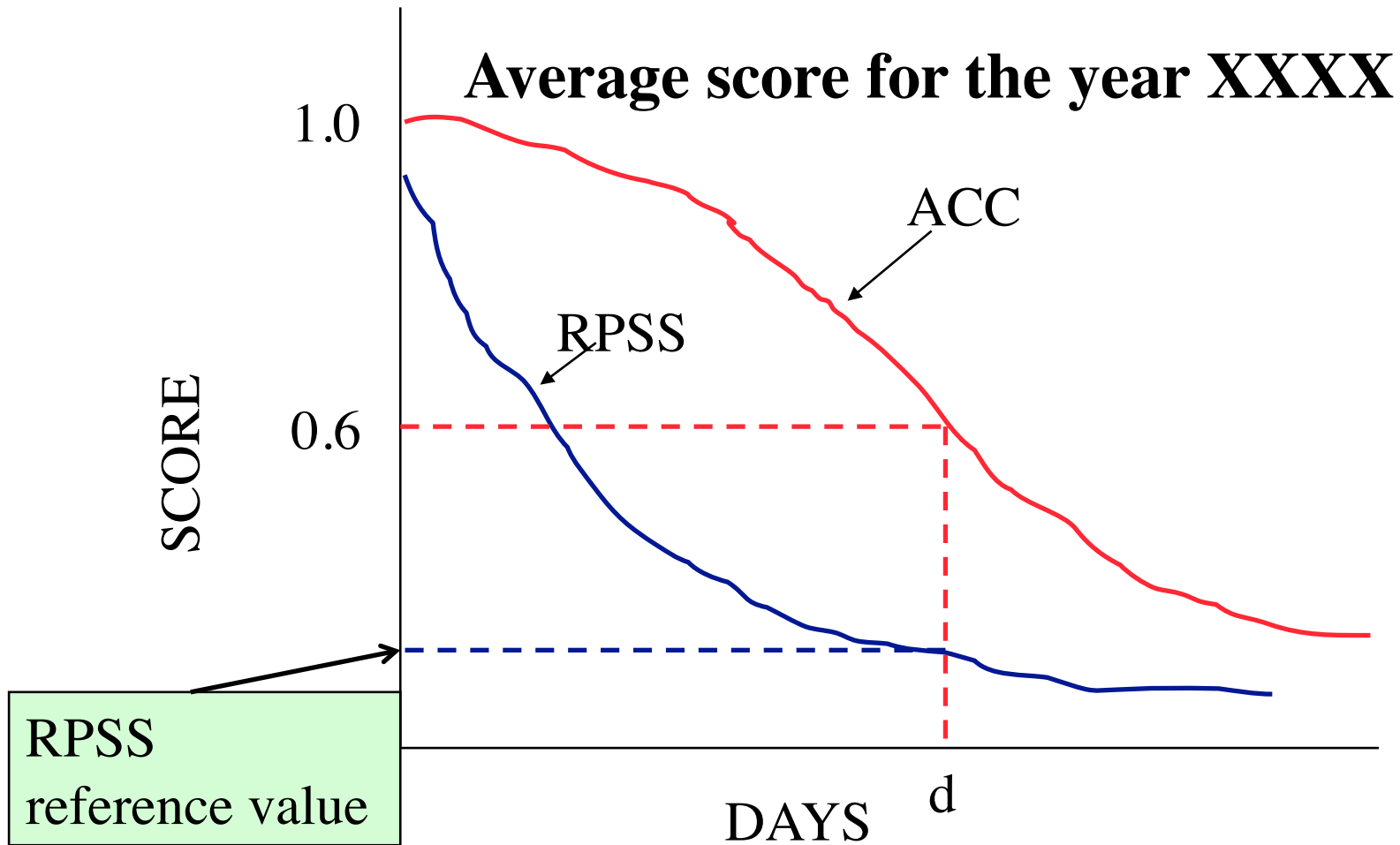
Target: Is a higher resolution model worth the resources spent on it?

RPSS, z500hPa, N.-Hem. (20N-90N), Control

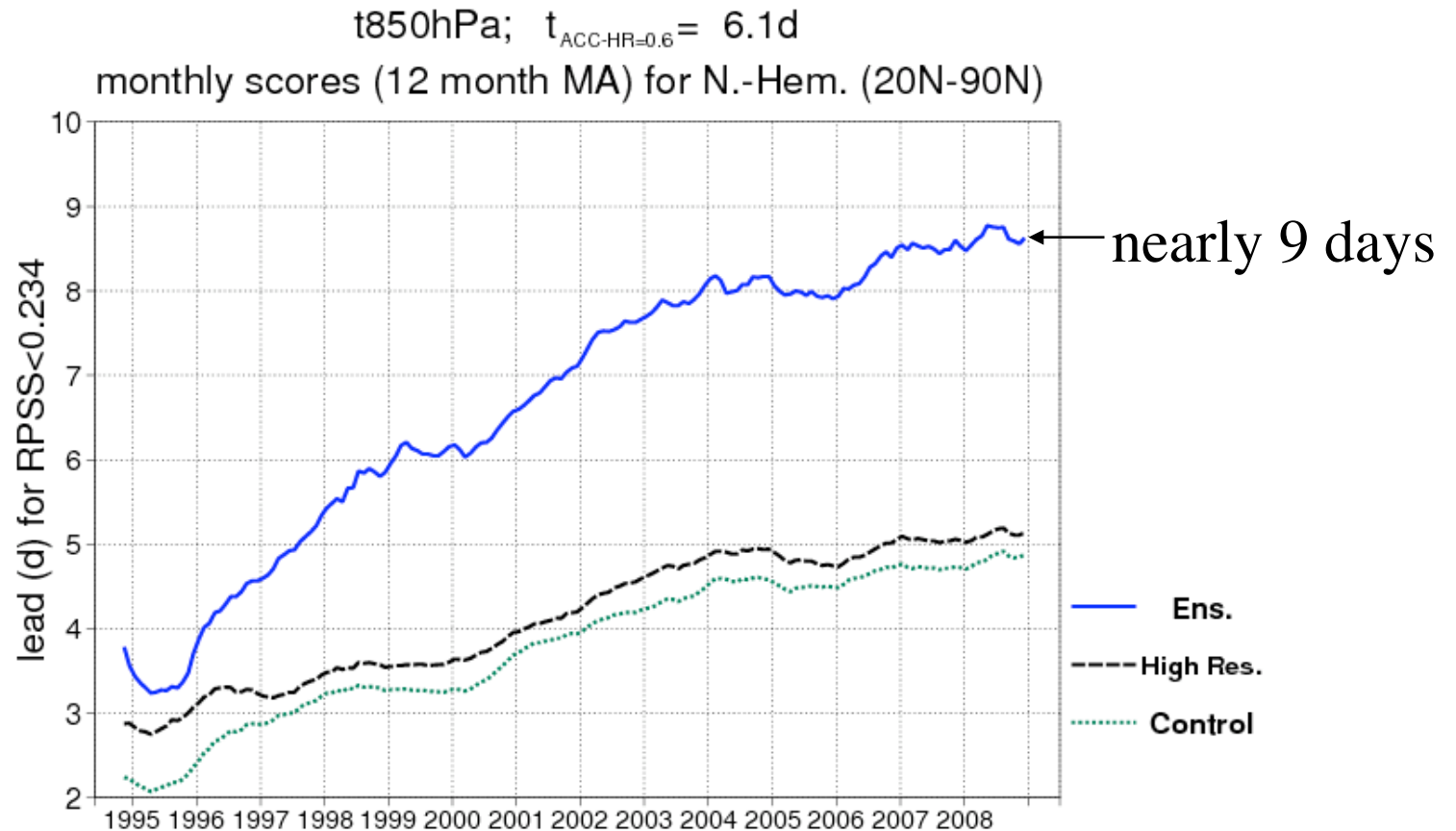


Looking at the long term improvements

The administrator's perspective



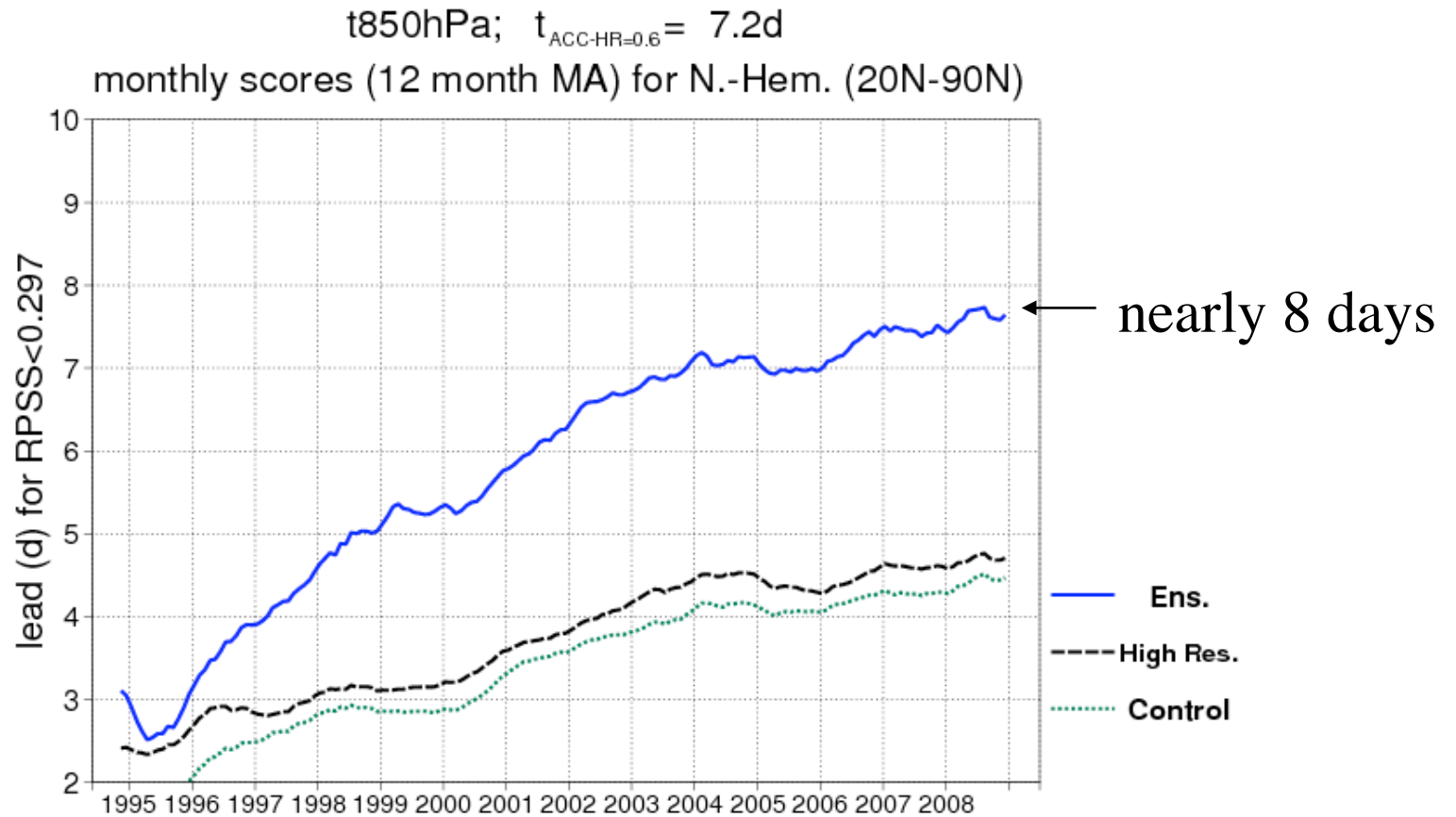
Looking at the long term improvements



Reference date 1999

Looking at the long term improvements

The administrator's perspective



Reference date 2006

Looking at seasonal improvements

The modeler/forecaster

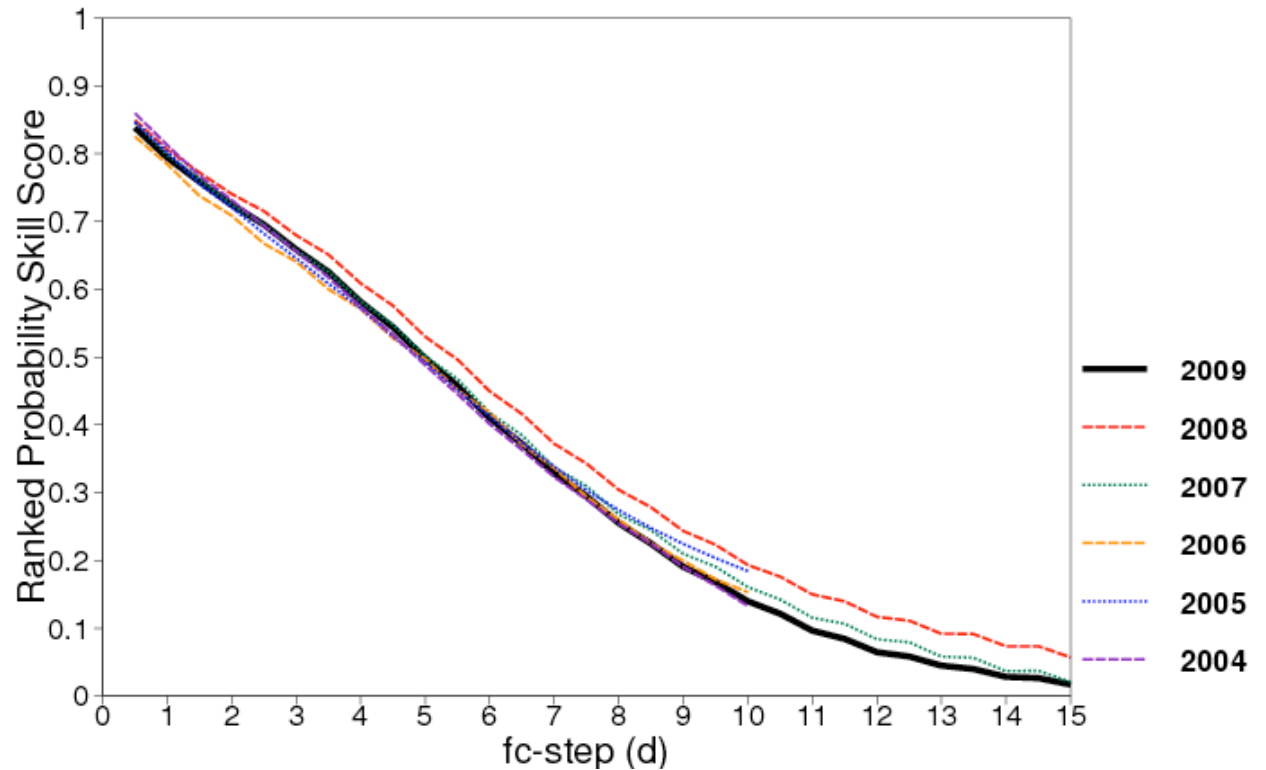
This spring was better (or worse) than other springs.
Why?

Model changes

Atmospheric flow

Sample size changes

t at 850hPa
10 categories (Quan), area n.hem
MAM



Measuring accuracy

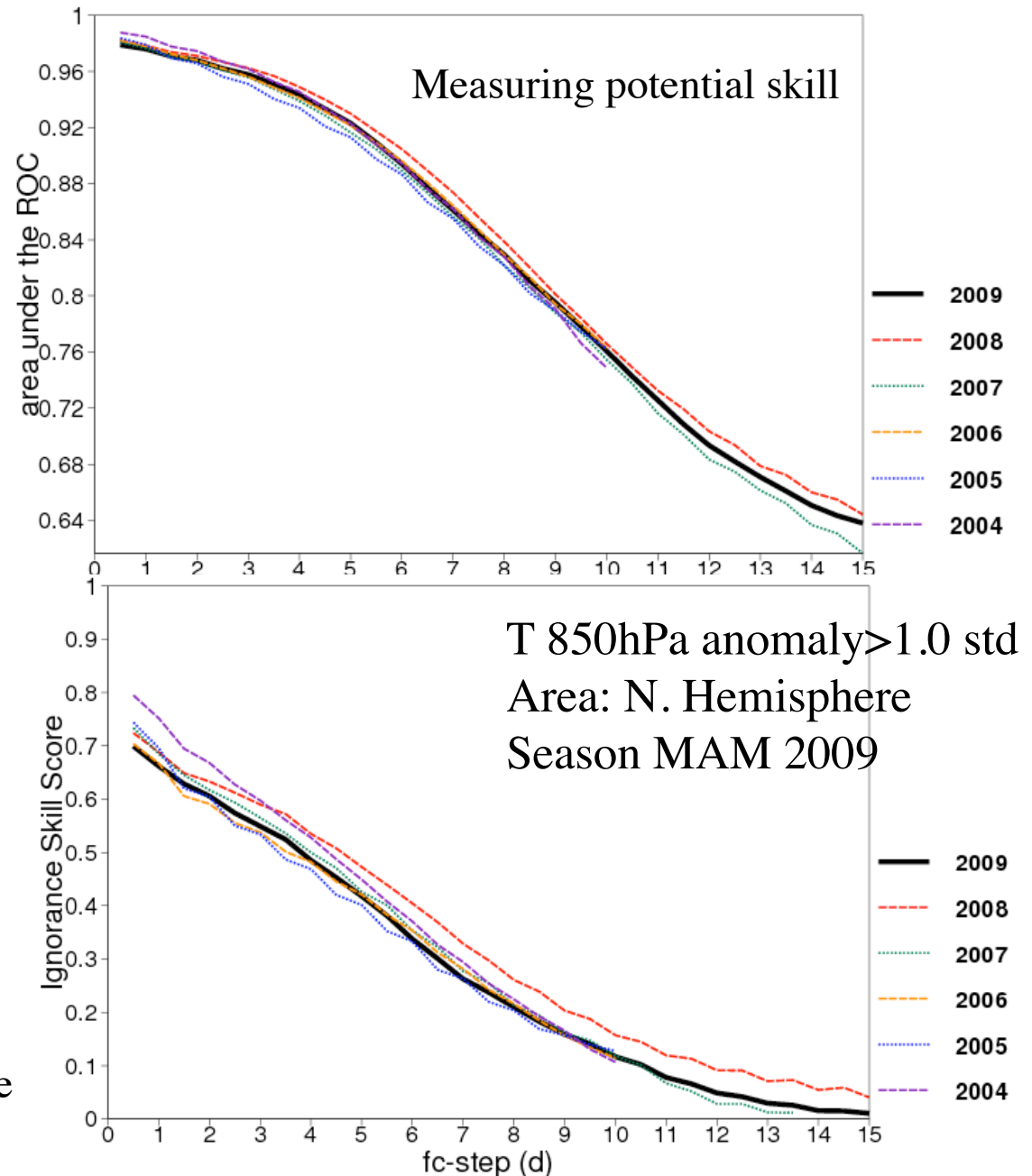
Looking at seasonal improvements

The modeler/forecaster

This spring was better (or worse) than other springs.
Why?

- Model changes
- Atmospheric flow
- Sample size changes

Measuring value



Comparing thresholds and time improvements

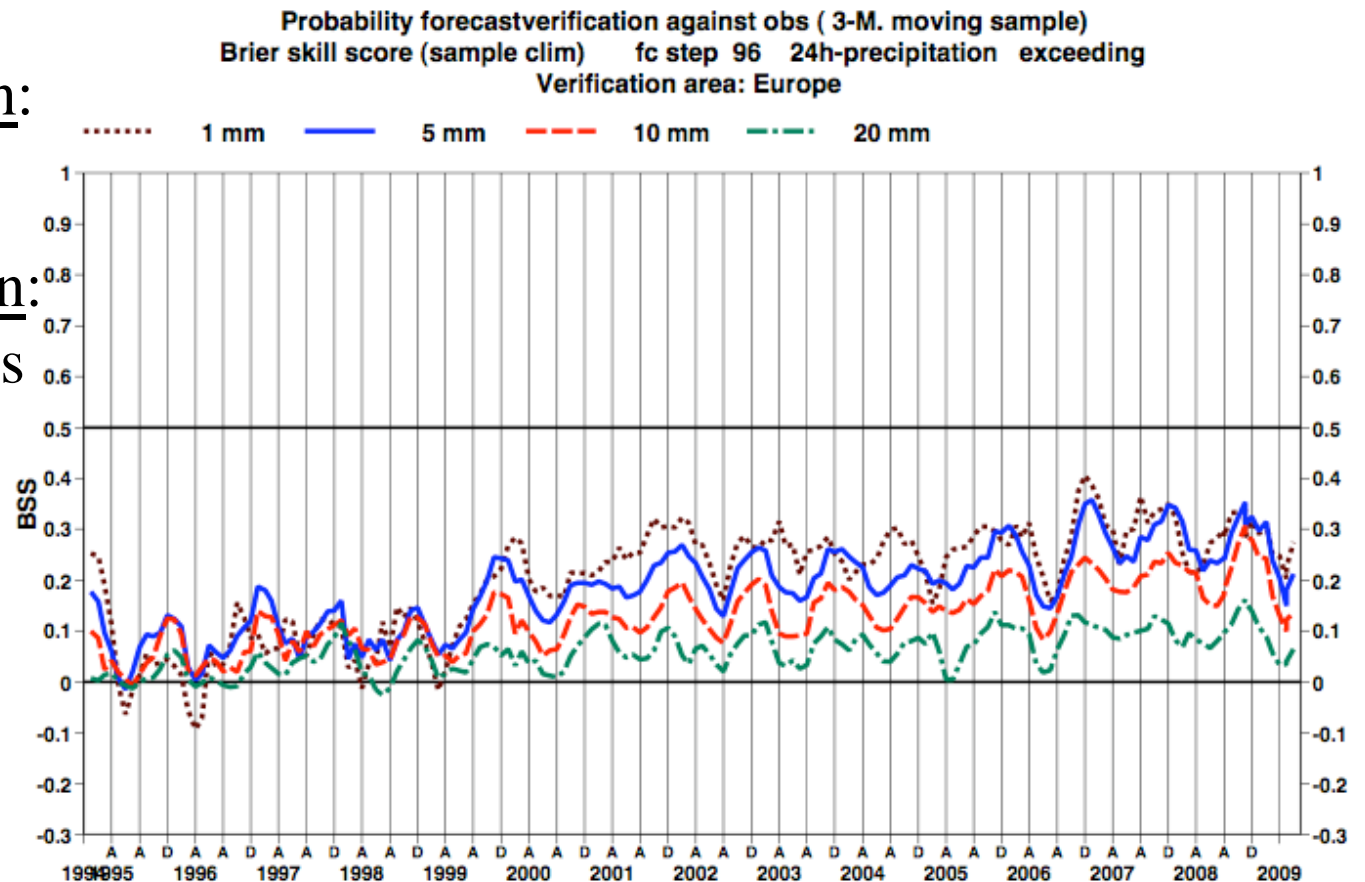
Modeler's question:

Is there any trend?

End user's question:

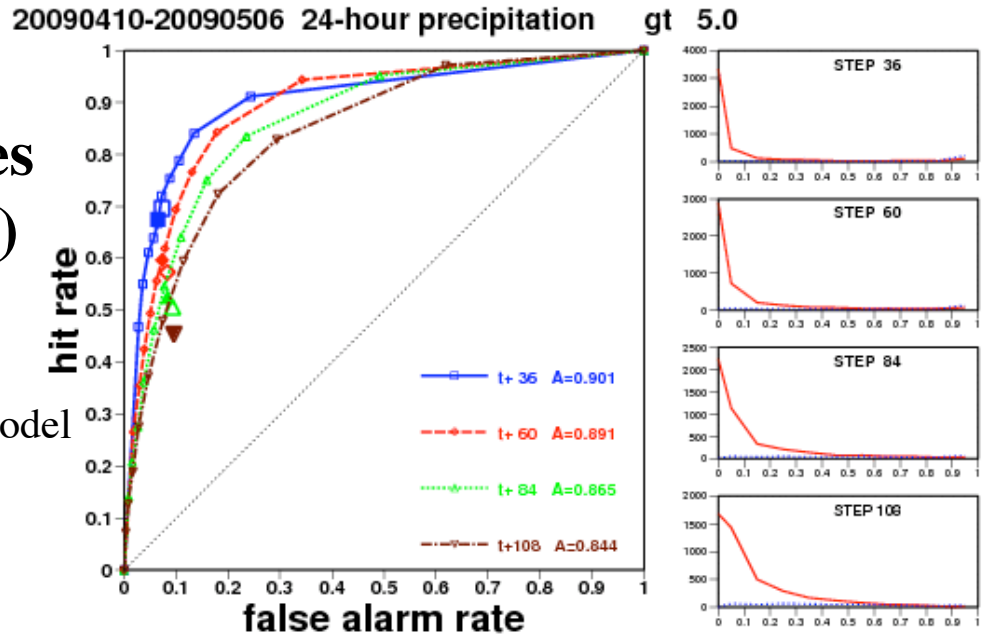
How much better is the ensemble than climatology?

Measuring total error



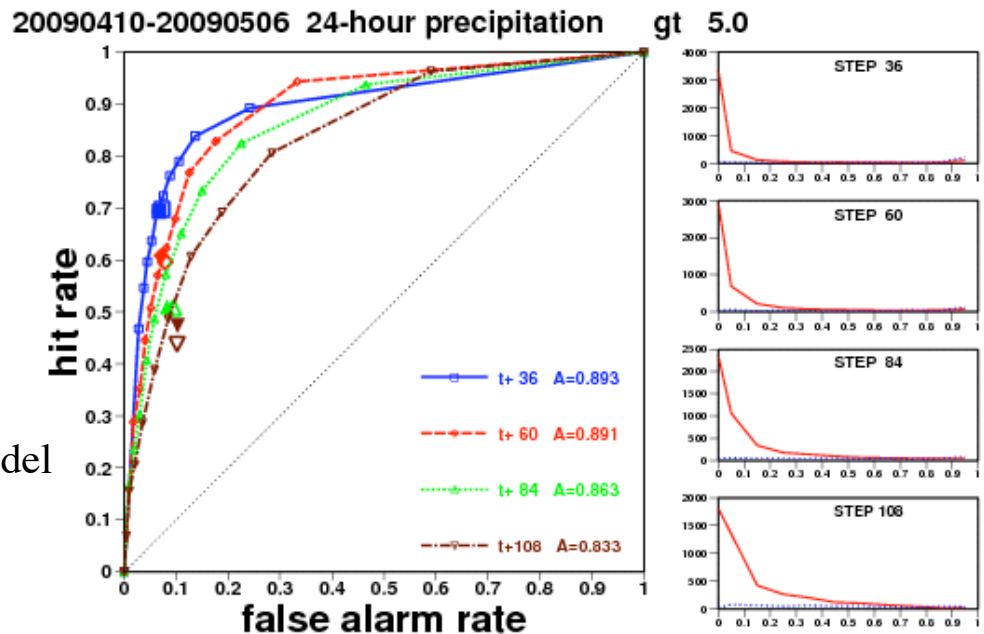
Comparing experimental suites (different model specifications)

Operational model



The modeler's question:
Is the new model formulation
better than the old one in
terms of discrimination?

Experimental model



Measuring potential skill

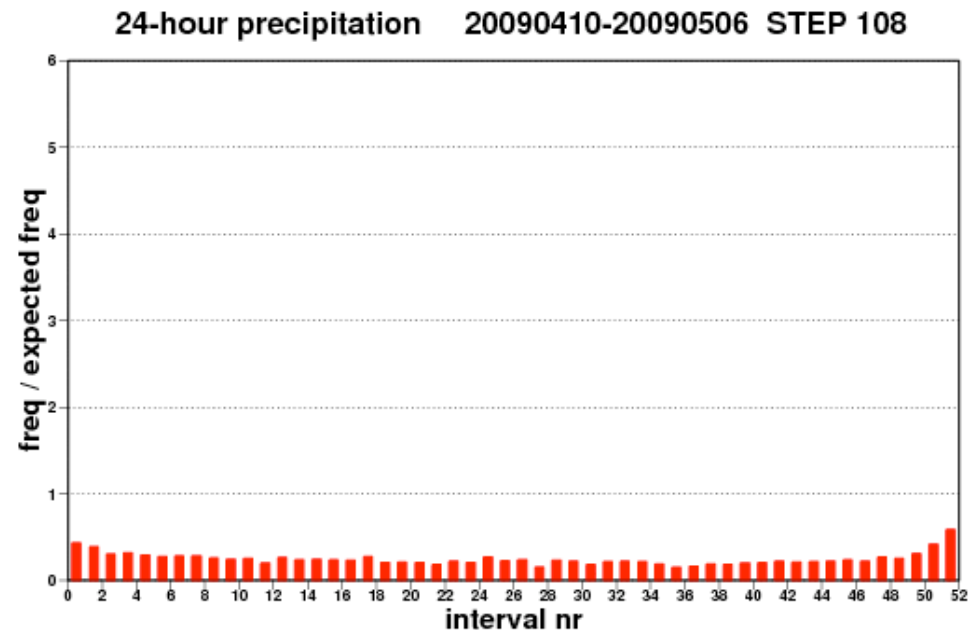
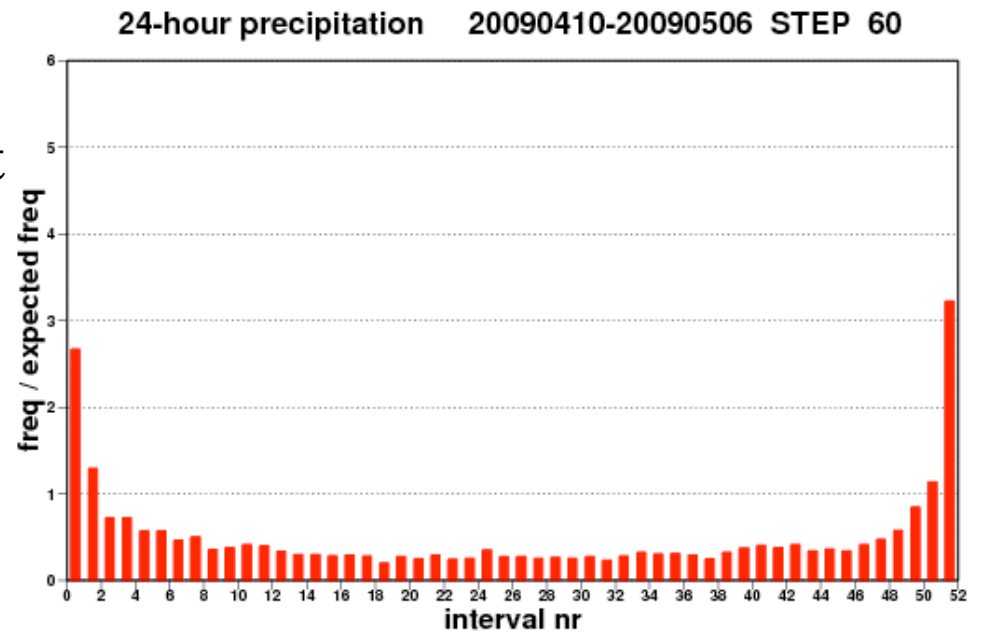
Talagrand diagram: a snapshot

At $t+60$ U-shape --> the spread is too small to represent uncertainty

At $t+108$ flat --> the spread is about right to represent uncertainty

The modeler: Should the spread be increased?

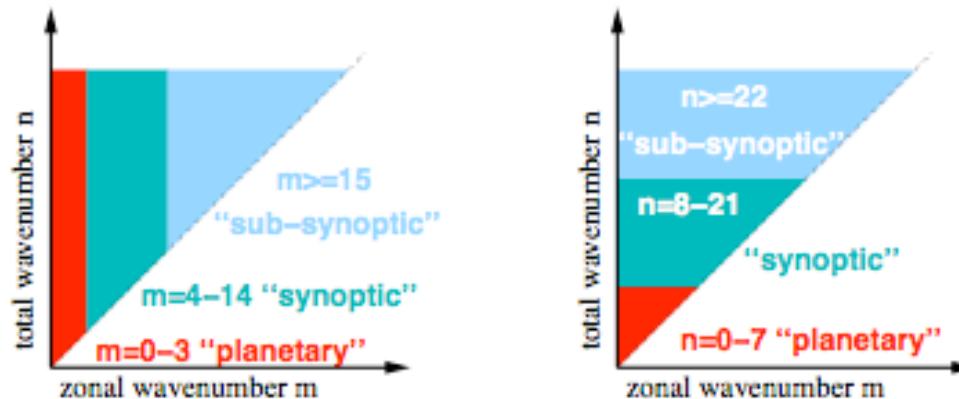
The forecaster: should I use the system for time ranges $< t + 60$?



scale-dependent verification: methodology

$$Z(\lambda, \phi) = \sum_{m=-M}^{m=M} \sum_{n=|m|}^M a_{m,n} Y_{m,n}(\lambda, \phi)$$

$Y_{m,n}(\lambda, \phi) = P_{m,n}(\sin \phi) e^{im\lambda}$, where λ, ϕ denote longitude and latitude, resp.

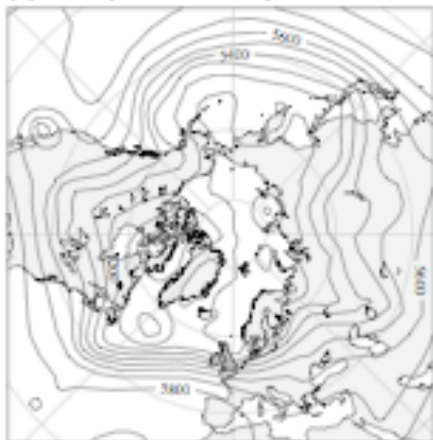


- synoptic band corresponds to **wavelength 2000–7000 km** for the zonal wavenumber filter at 45 and to wavelength **2000–5000 km** for the total wavenumber filter
- lower wavenumber limit of the “synoptic” bands is motivated by wavenumber spectra of low and high-frequency filtered fields (geopotential and wind)

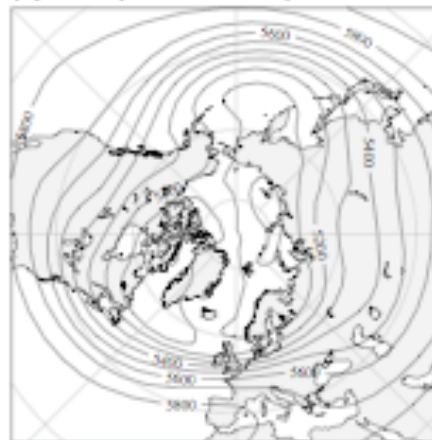
Jung and Leutbecher, QJ 2008

Filtering by zonal wavenumber: 18 Jan '07, 12 UTC

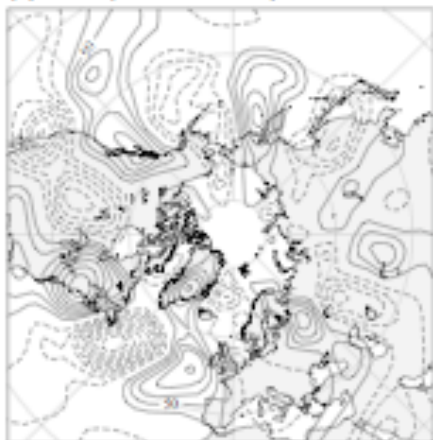
(a) Z500 (20070118 12z): zwn k=0-159



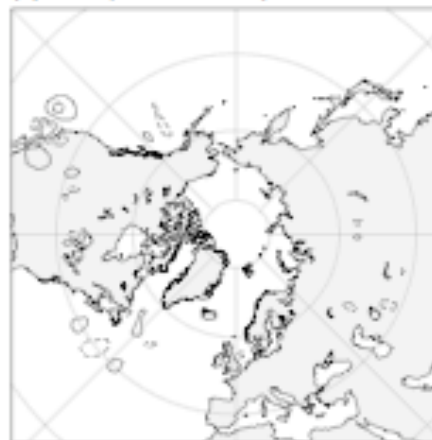
(b) Z500 (20070118 12z): zwn k=0-3



(c) Z500 (20070118 12z): zwn k=4-14

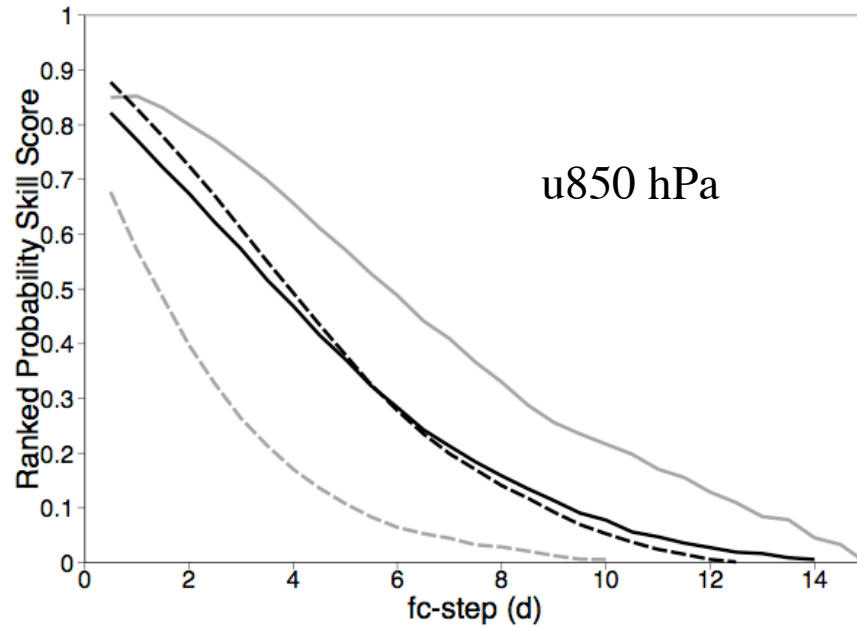
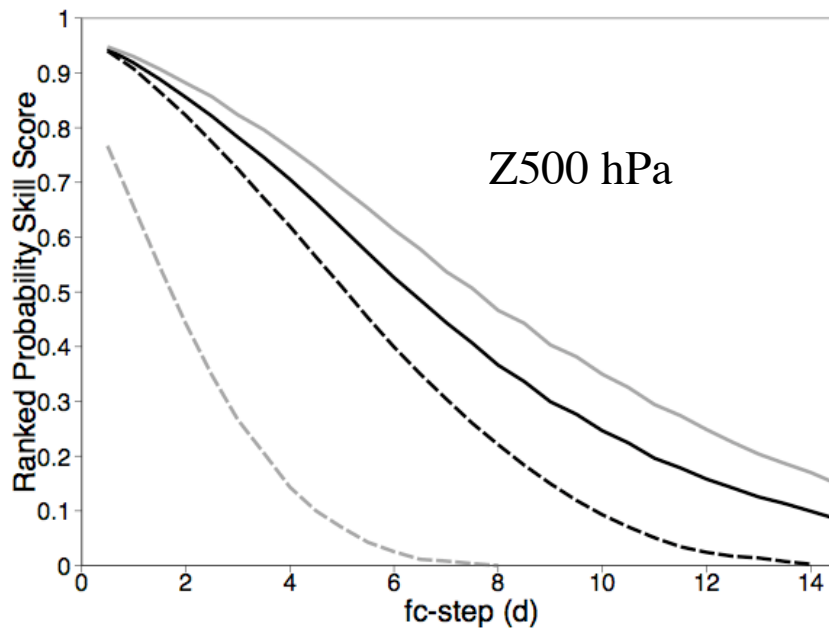


(d) Z500 (20070118 12z): zwn k=15-159

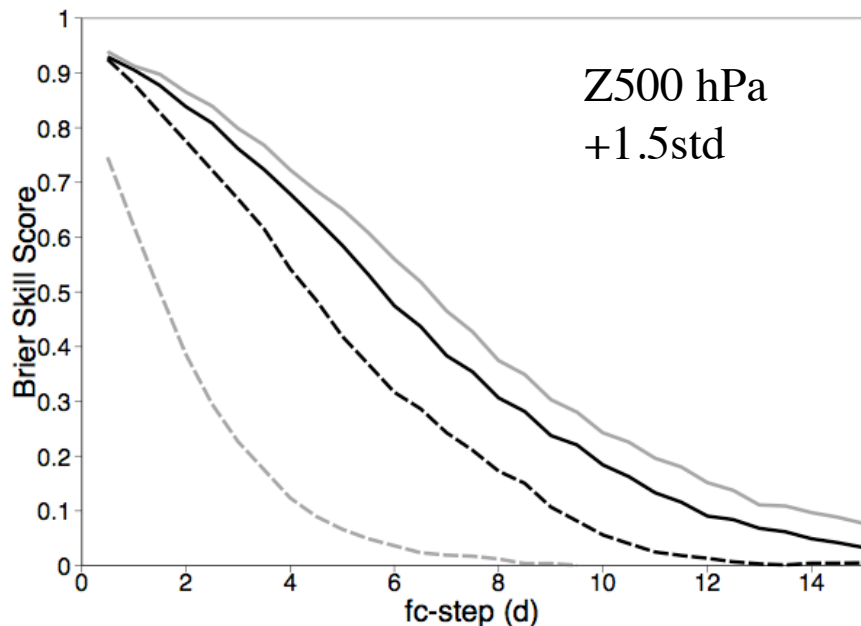


contour
intervals
(a) 100 m
(b) 100 m
(c) 50 m
(d) 25 m

Jung and Leutbecher, QJ 2008



— UF
 — M0-3
 - - M4-14
 - - M15-159



— UF
 — M0-3
 - - M4-14
 - - M15-159

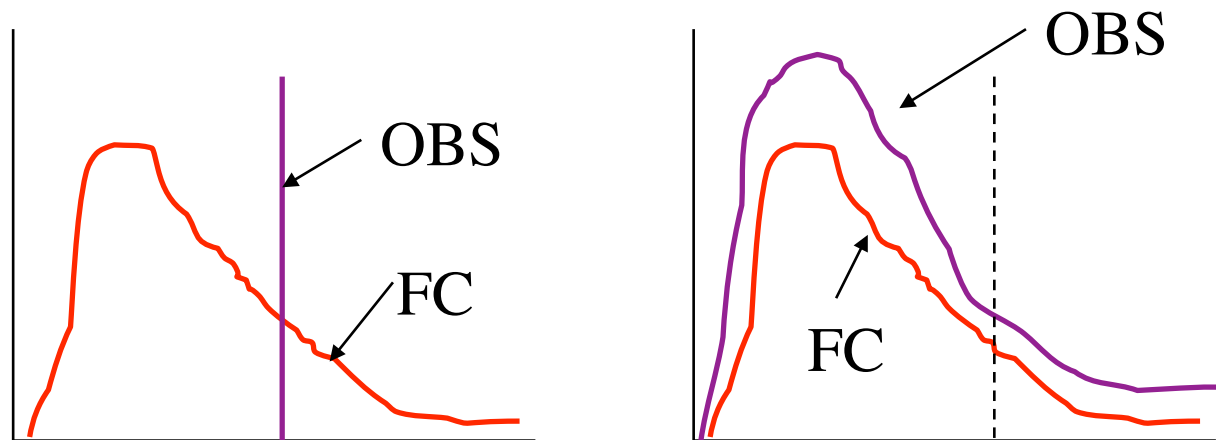
Outcome:

1. Planetary scales
2. Synoptic scales
3. Subsynoptic scales

Jung and Leutbecher, QJ 2008

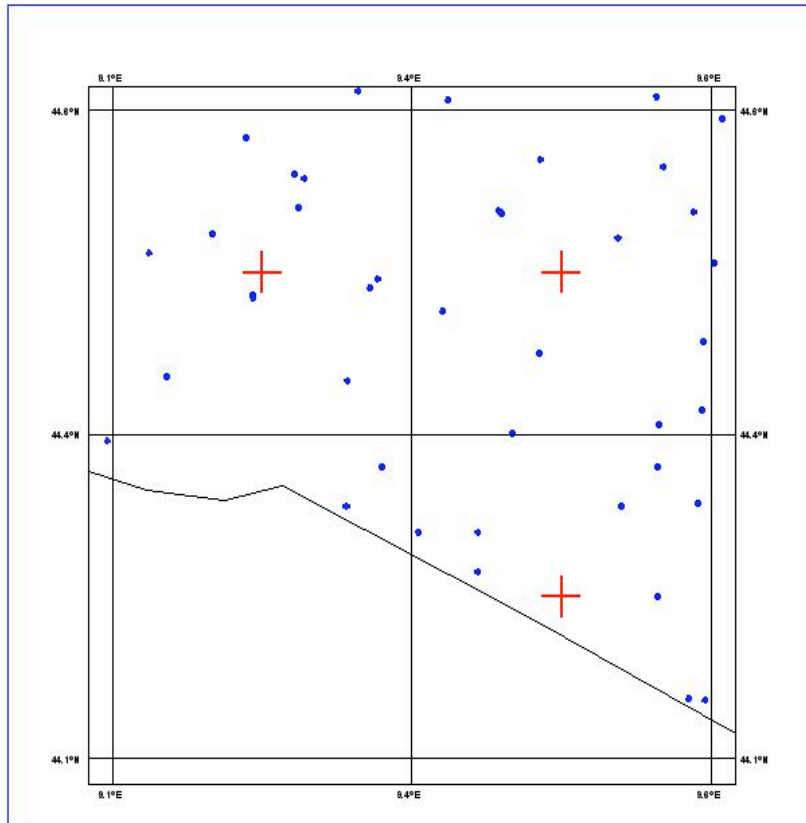
Introducing observation uncertainty

- Following “Observational probability”
Candille&Talagrand, 2008
- Using up-scaled observations to build obs
PDF
- How does the ensemble performance differ?



Carlos Santos, AEMET, Spain

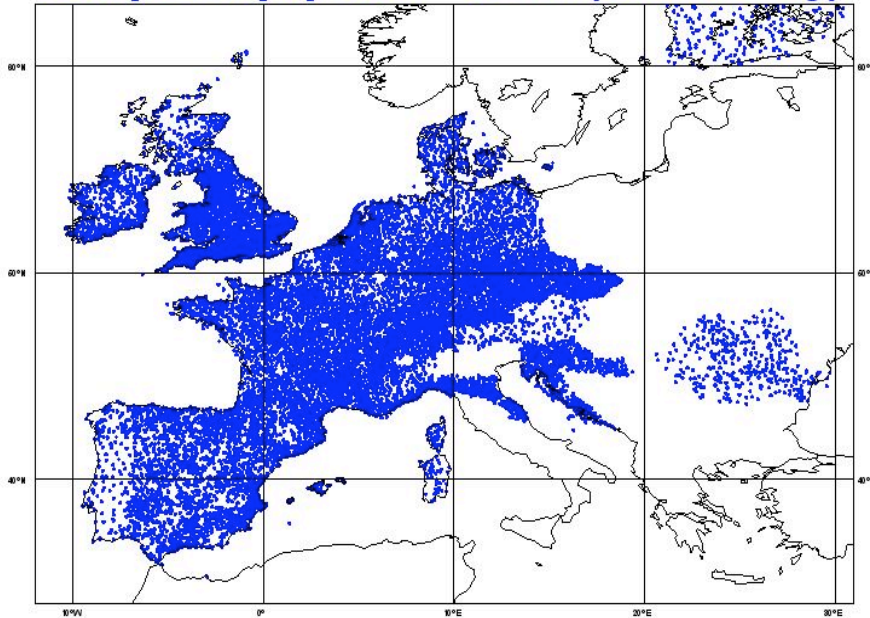
Up-scaling method



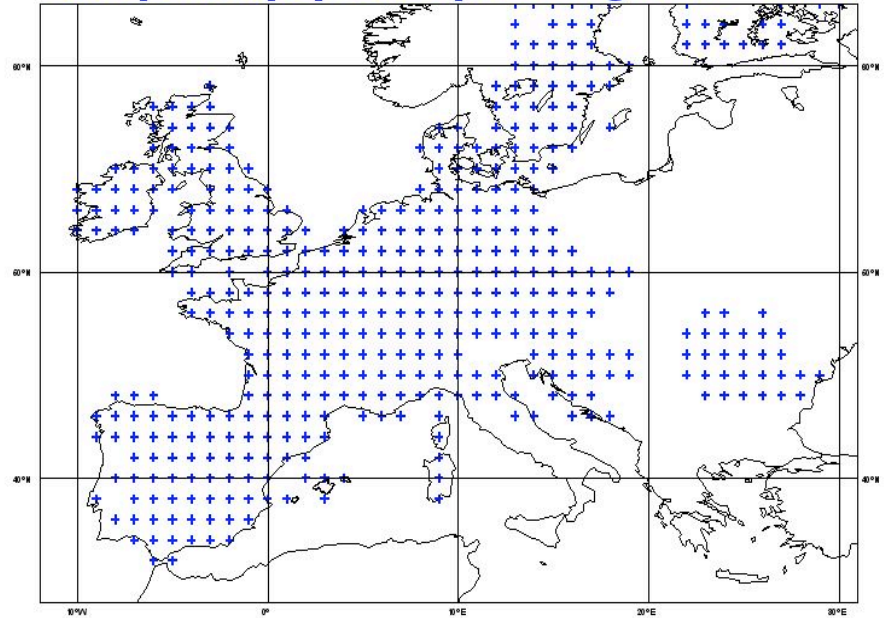
- Up-scaling Europe HR data
- Two up-scaling box sizes:
 - $1^\circ \times 1^\circ$
- Minimum number of obs:
 - #obs < 5 \rightarrow grid point rejected
 - #obs ≥ 5 \rightarrow grid point OK
- Two different representations of “truth”:
 - Average
 - Quantiles 10, 25, 50, 75, 90

Up-scaling method

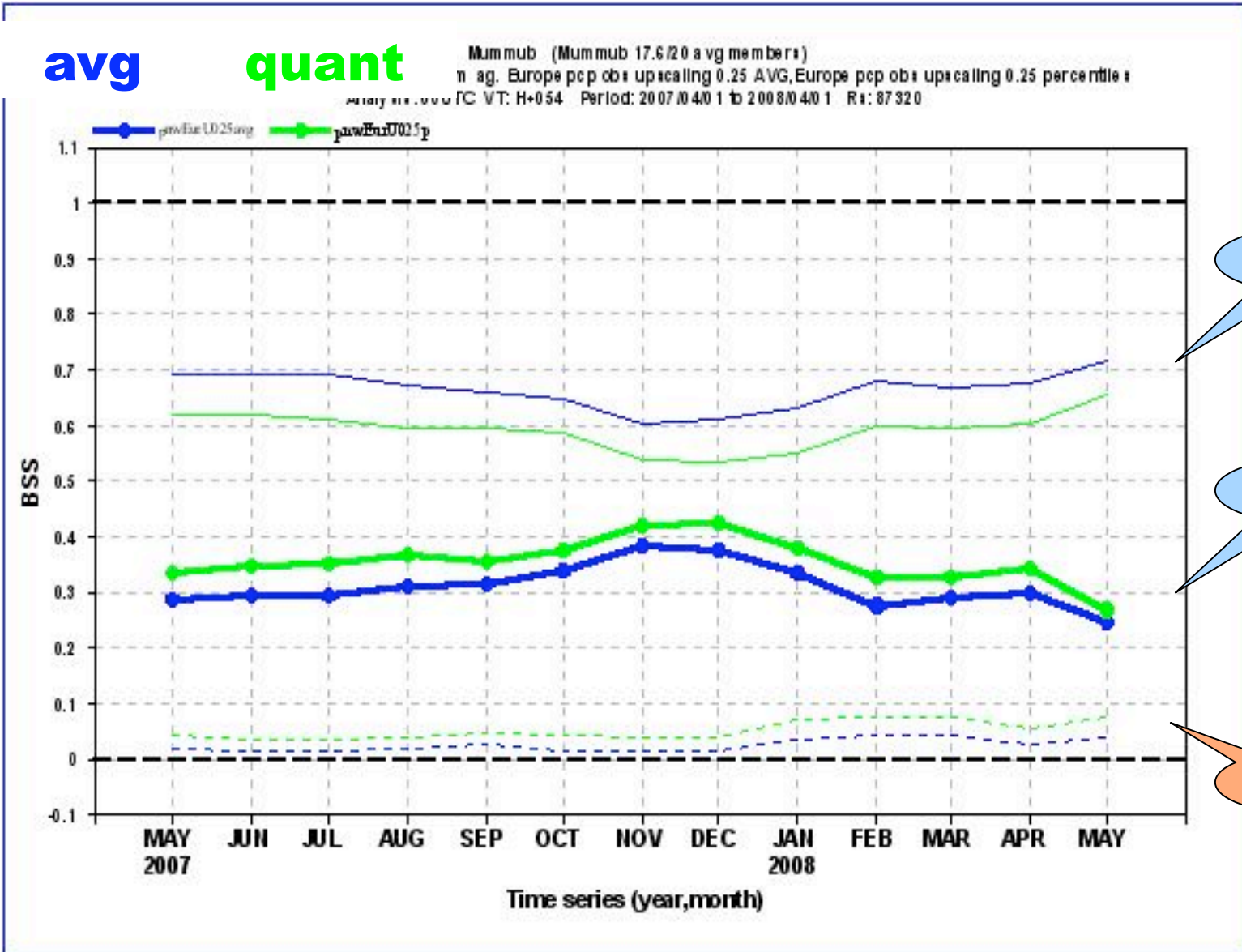
Europe HR pcp obs ~ 16000 (random day)



Europe HR pcp obs up-scaling 1.00 ~ 500



BSS generally ↑



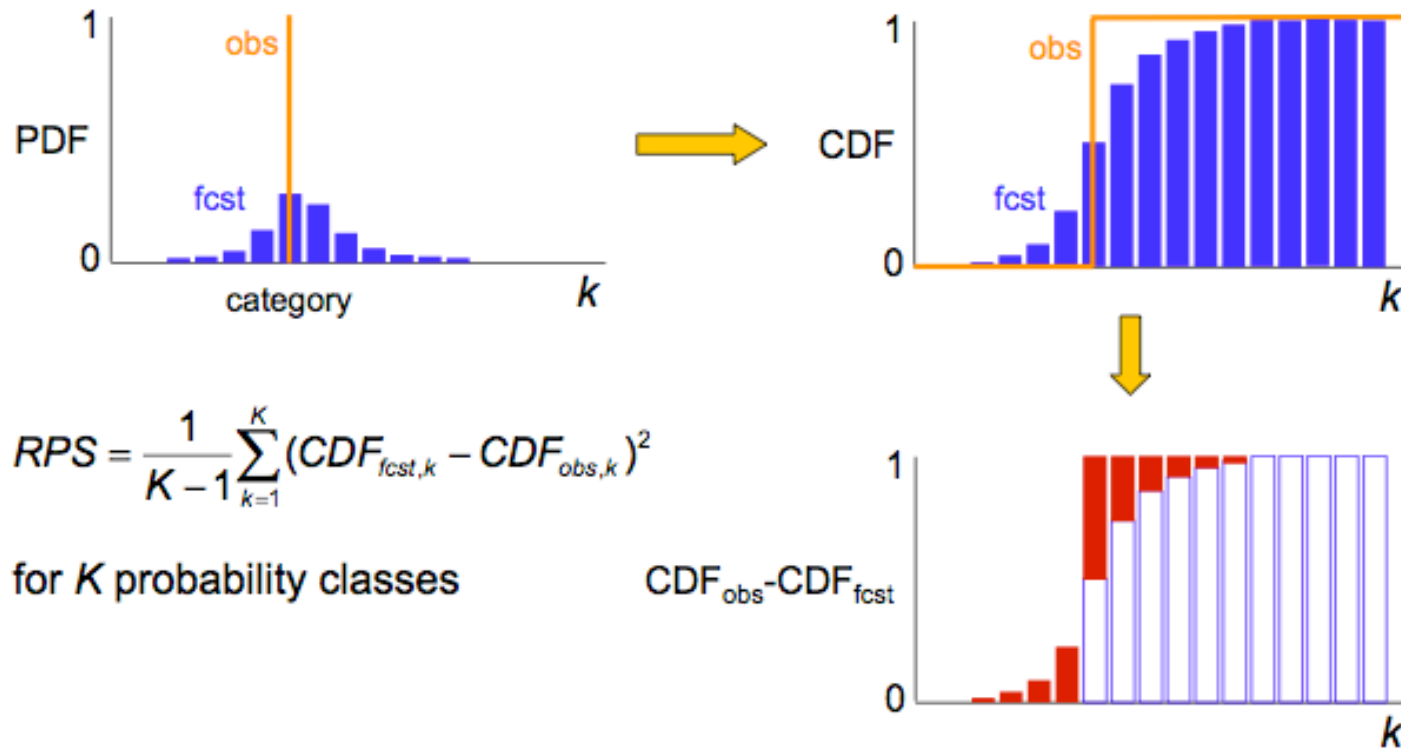
Conclusions

- **Each metric gives information on a different aspect of forecast performance**
- **The verification question**
- **The users**
- **Model performance: case studies and/or long term trends. The implications for the verification software/packages**

My Questions:

- **Can we suggest a list of scores that are enough to assess a model for a selection of specific users?**
- **Are metrics parameter-dependent?**

Ranked Probability Score



Discrimination

A good forecast should be able to discriminate between event and non-event

