



# New **ECMWF** Headline Verification Measures

**Pertti Nurmi**

*Meteorological Research & Development*

*Finnish Meteorological Institute*

( [pertti.nurmi@fmi.fi](mailto:pertti.nurmi@fmi.fi) )

**Convenor of ECMWF/TAC Subgroup on Verification Measures**



*Acknowledgement to all 12 Subgroup members and ECMWF staff*



# Introduction

- ✓ **Verification Expert Group established by 39.TAC, Oct 2008**
- ✓ **Participants**
  1. Alexander Kann – Austria
  2. Marielle Amodei – France
  3. Martin Göber – Germany
  4. Gabriella Csima – Hungary
  5. James Hamilton – Ireland
  6. Kees Kok – Netherlands
  7. Carlos Santos – Spain
  8. Marion Mittermeier – U.K.
  9. Frank Saunders – U.K.
  10. Clive Wilson (EUMETNET)
  11. Chris Ferro (invited expert)
  12. David Stephenson (invited expert)
  13. Pertti Nurmi – Finland (Convenor)
- ✓ **Active participation and preparatory work by ECMWF staff**
- ✓ **Four meetings, Mar 2009 ... Sep 2010 >> Proposal Oct 2010**



# Terms of Reference

- Investigate and provide recommendations on verification measures appropriate to **monitor ECMWF's long-term progress** in medium-range forecasting, with particular emphasis on **early warnings of severe weather**
- ⇒ Recommend **headline measures** suitable to complement those of current ECMWF Strategy (valid until end 2010)  
*(i.e anomaly correlation of Z500 for the deterministic forecast and probabilistic scores of T850 for the EPS)*
  - ✓ Recommend verification procedures to **aid forecasters' decision making**
  - ✓ Recommend measures suitable for verification of forecasts of **high-impact events**
  - ✓ Identify **requirements for observational data** necessary for verification



## High-impact weather

- ✓ **Comprehensive review of available verification scores for high-impact (severe) weather was performed by the Group, including presently on-going developments**
- ✓ **The Group identified two major issues:**
  1. **Lack of observations at sufficient temporal and spatial resolution**
  2. **Lack of fundamental research into related verification in meteorological services and universities**
- ✓ **The Group identified a set of properties that verification scores (especially for extreme events) should possess**
  - **No currently available measures satisfy these requirements**
- ⇒ **Substantial research is needed to develop suitable verification scores - both for deterministic and probabilistic forecasts**



# Headline measures

- **To consider**: When was the latest, generally accepted verification measure “invented” ??? ⇔ **NOT** a trivial task !!!
- **From** upper air verification >>> **To** surface weather verification
  - ✓ Subgroup reviewed surface weather variables and scores to supplement present ECMWF upper-air verification
- ⇒ **... Precipitation accumulated over 24 hours**
  - ✓ Is **\*THE\*** most significant weather event relevant for end-users
  - ✓ Is a **DMO** parameter (*e.g. T2m & 10m wind are post-processed*)
  - ✓ **24-hour accumulation** mitigates representativeness issues
  - ✓ **Both deterministic and EPS output** need to be verified
  - ✓ **Verification to be done against SYNOPs**, applying basic QC, applying nearest grid point approach  
(*WMO JWGFVR recommendation*)



# Headline measures

- a) Primary scores for synoptic-scale upper-air fields ⇔ As before**
- ⇒ Deterministic forecast - Anomaly correlation for 500 hPa height
  - ⇒ EPS - Probabilistic skill for 850 hPa temperature
  - ✓ Verification against analyses
  - ✓ Well-established, standard scores, meet WMO requirements
- b) Supplementary scores for surface weather and severe weather**
- ⇒ Surface weather: Deterministic & EPS - 24-hour precipitation
  - ⇒ Severe weather: Tropical Cyclone position error - EFI skill
  - ✓ Verification against observations (*additional uncertainties*)
  - ✓ No WMO-designated scores for these parameters !
    - ⇔ Potential future action (?)



# Headline measures

## Issues:

- ✓ Long-term trends in performance will be monitored by considering: **The number of days for which forecast skill remains above a pre-specified threshold**  
⇔ **Intuitive interpretation**
- ✓ Each individual headline score should be **presented separately**  
⇔ **Scores were chosen to indicate different aspects of forecast performance progress at a range of spatial and temporal scales**
- ✓ **Direct comparison of the different scores or their combination into a single index is \*NOT\* appropriate**



# Headline measures

- **Strategy, 2011 – 2020:**

*To follow up long-term performance of the ECMWF...*

**a) Keep the 2 existing primary headline measures**

i. **Anomaly correlation, ACC, for the 500 hPa geopotential height for the *deterministic forecasts***

ii. **Continuous ranked probability skill score, CRPSS, for the 850 hPa temperature for the *probabilistic forecasts***

**b) Introduce 4 supplementary headline measures**

Proposal

Proposal





# Headline measures

Long-term trends are monitored by:  
The number of days for which the forecast skill  
remains above a pre-specified threshold

## a) Keep the 2 existing primary headline measures

- i. Anomaly correlation, ACC, for the 500 hPa geopotential height for the *deterministic forecasts*

→ An appropriate threshold should be **80%**  
currently achieved at ~ Day 6

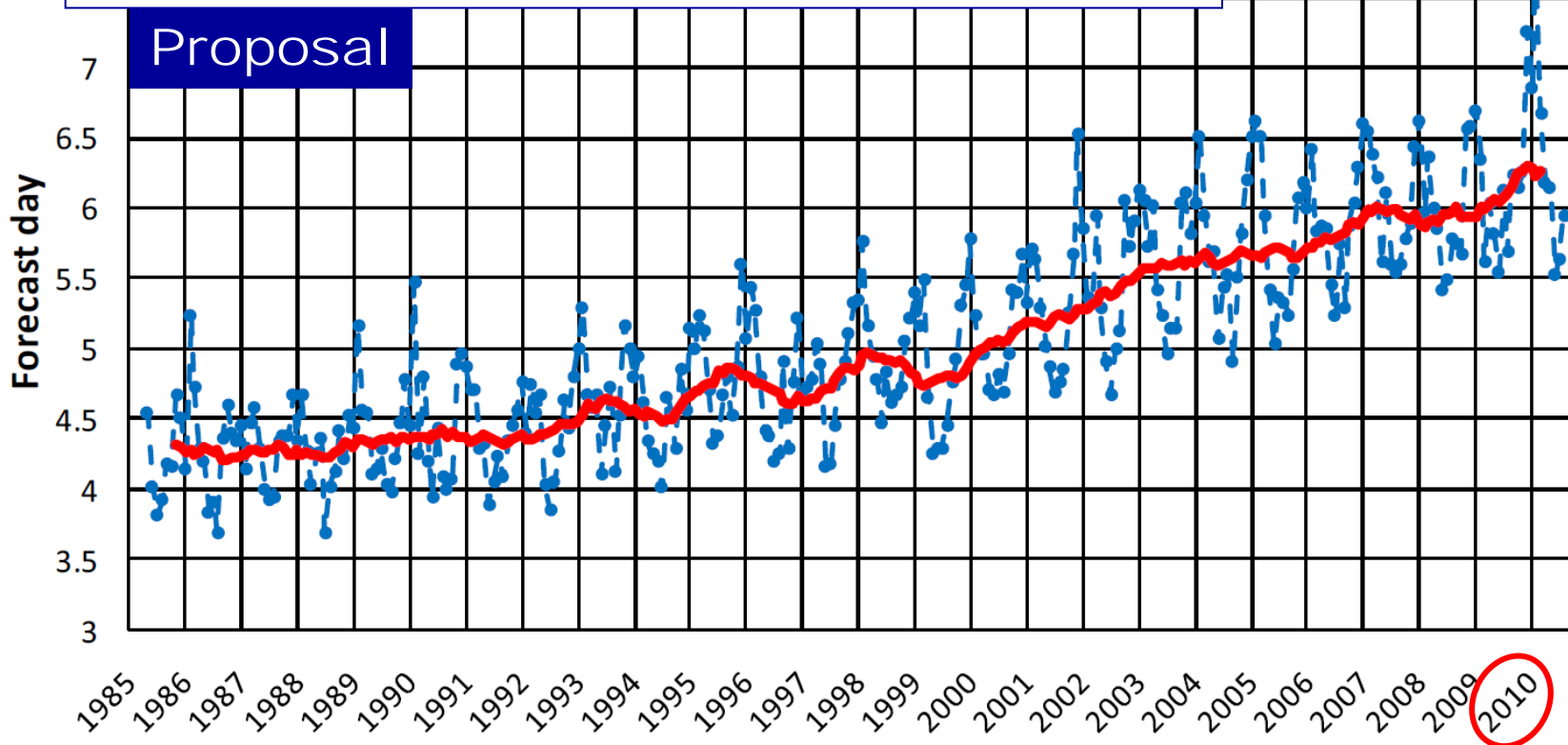
- ii. Continuous ranked probability skill score, CRPSS, for the 850 hPa temperature for the *probabilistic forecasts*

→ An appropriate threshold should be **25%**  
currently achieved at ~ Day 8



# Headline measures

ACC - Z500 hPa: Remains above 80%

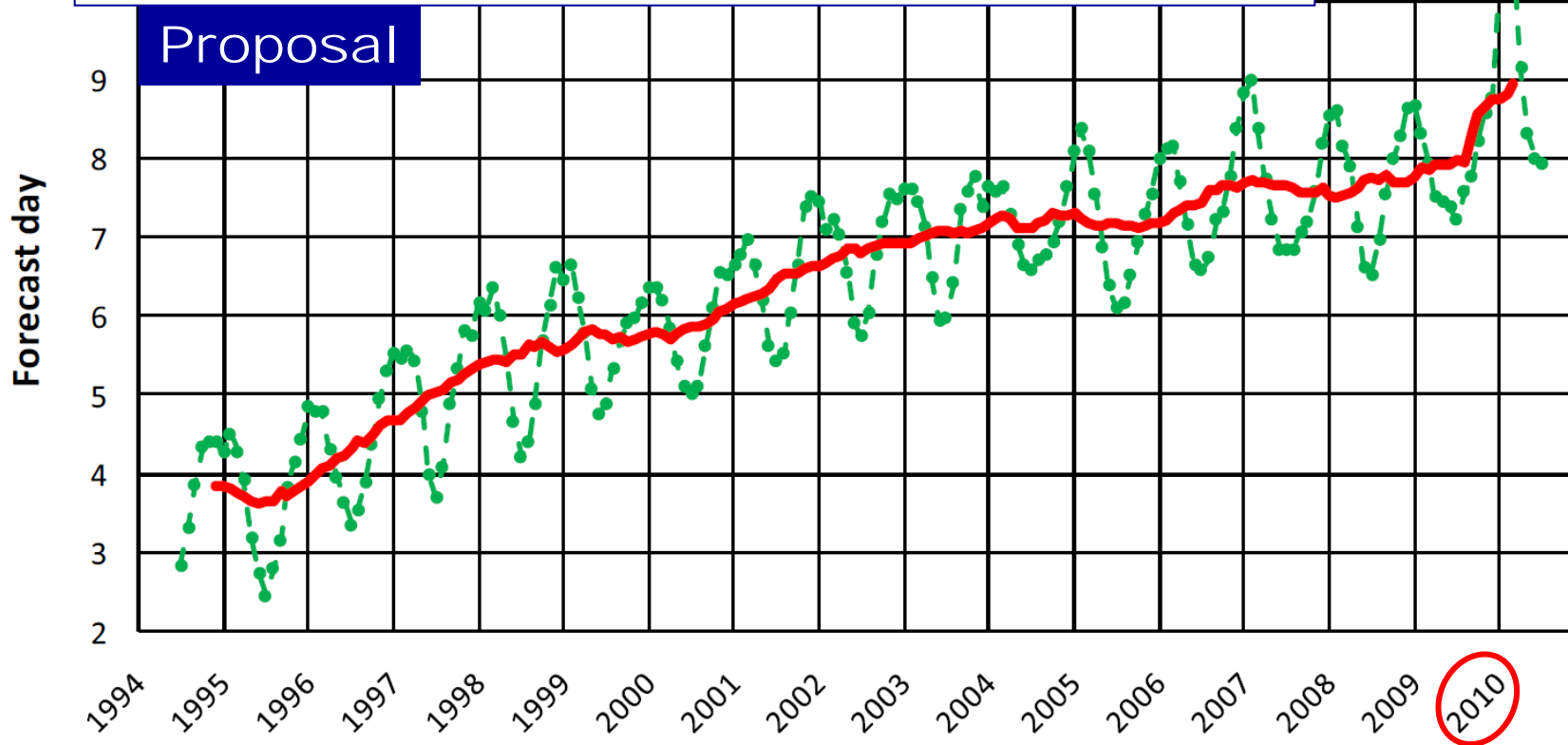


Primary headline score for **deterministic** forecasts. Each curve shows the number of days for which the monthly mean (blue line) or 12-month mean centred on that month (red line) deterministic forecast skill remains above a specified threshold for 500 hPa height operational forecasts over the extra-tropical northern hemisphere. The threshold is chosen to reflect the **deterministic** forecast skill that is achieved at day 6-6.5 at the beginning of the strategy period.



# Headline measures

CRPSS - T850 hPa: Remains above 25%



Primary headline score for **probabilistic** forecasts. Each curve shows the number of days for which the centered 3-month mean (green line) or 12-month mean (red line) probabilistic forecast skill remains above a specified threshold for 850 hPa temperature forecasts from the EPS over the extra-tropical northern hemisphere. The threshold is chosen to reflect the **probabilistic** forecast skill that is achieved at **day 8.5-9** at the beginning of the strategy period.



Long-term trends are monitored by:  
The number of days for which the forecast skill remains above a pre-specified threshold (**i and ii**)

**b) Introduce 4 supplementary headline measures**

- i. **New 1 - SEEPS score for the *deterministic forecasts* of 24-hour precipitation over the extra-tropics**
  
- ii. **CRPSS, for the *EPS probabilistic forecasts* of 24-hour precipitation over the extra-tropics**
  
- iii. **Severe weather: ROC Area for the *EFI for 10 m wind***
  
- iv. **Severe weather: Tropical cyclone position error for the *deterministic forecast***



Long-term trends are monitored by:  
The number of days for which the forecast skill remains above a pre-specified threshold (**i and ii**)

**b) Introduce 4 supplementary headline measures**

- i. **New 1 - SEEPS score for the *deterministic forecasts* of 24-hour precipitation over the extra-tropics**

→ An appropriate threshold should be **45%** currently achieved for the 24-hour period ~ 3.5 Days ahead

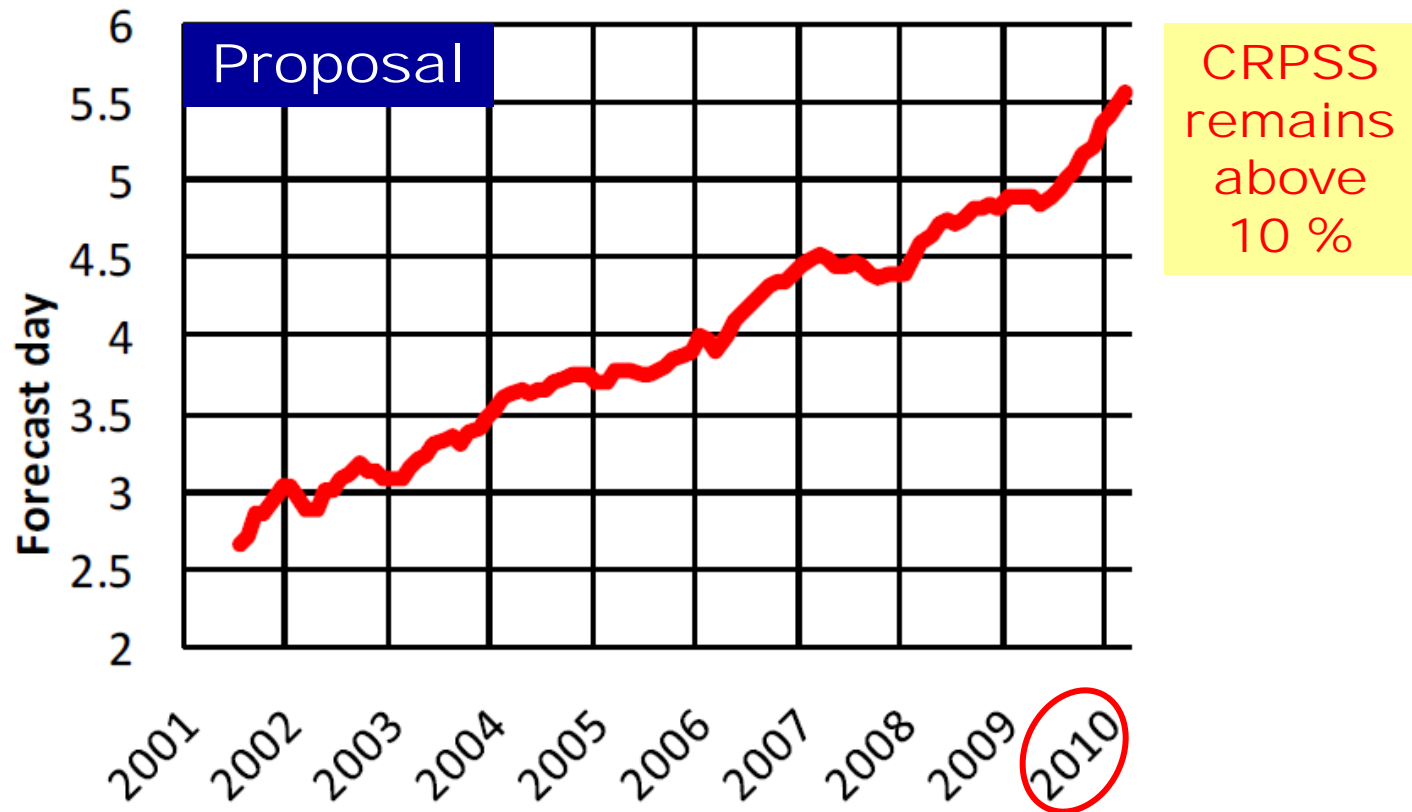
- ii. **CRPSS, for the *EPS probabilistic forecasts* of 24-hour precipitation over the extra-tropics**

→ An appropriate threshold should be **10%** currently achieved at ~ Day 5

- iii. **Severe weather: ROC Area for the *EFI for 10 m wind***

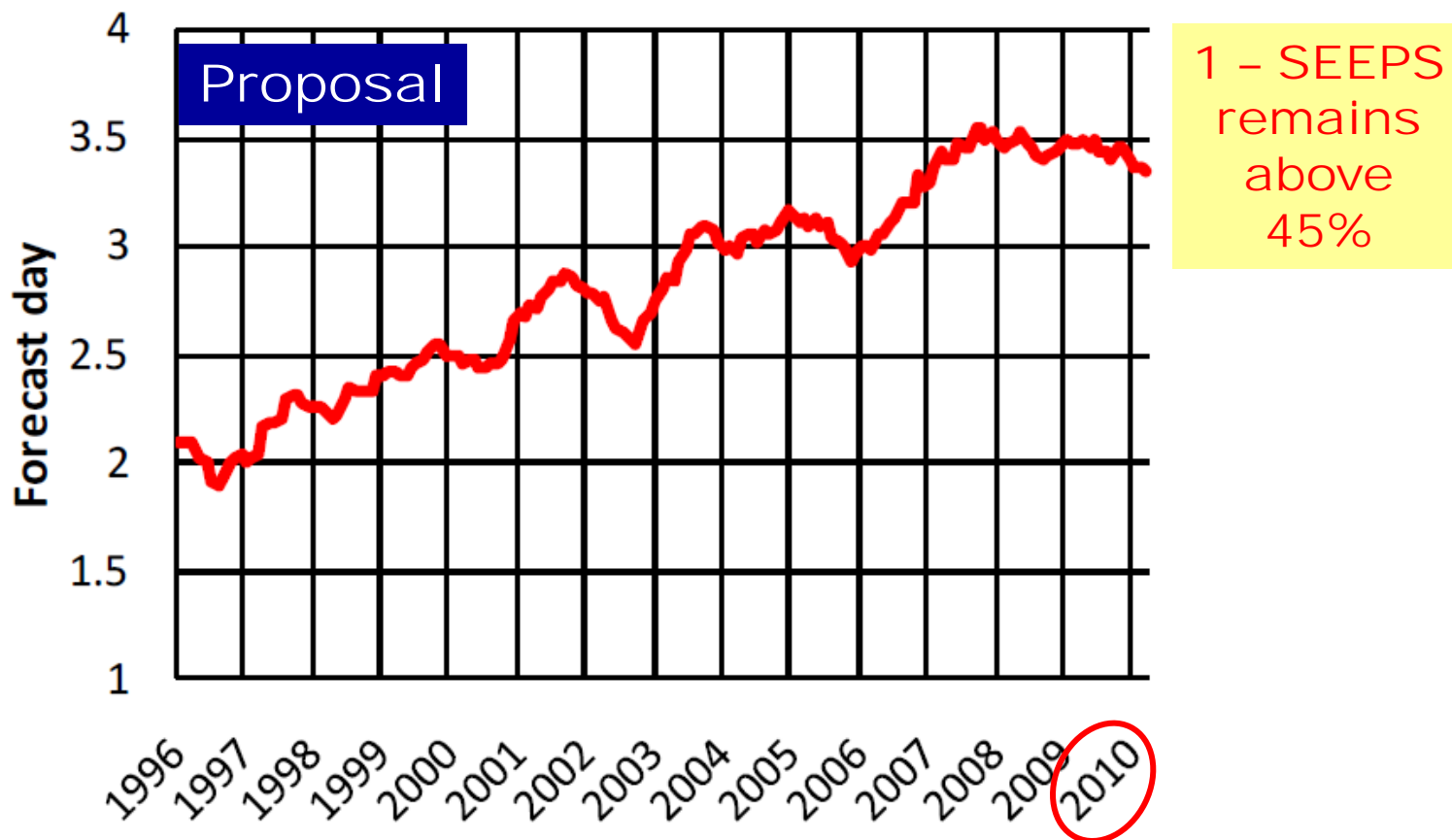
- iv. **Severe weather: Tropical cyclone position error for the *deterministic forecast***

# Supplementary headline measure (ii): CRPSS for 24-hr EPS Precipitation



Supplementary headline score for *probabilistic* precipitation forecasts. The curve shows the number of days for which the centered 12-month mean skill remains above a specified threshold for precipitation forecasts over the extra-tropics. The verification is for 24-hour total precipitation verifying against available synoptic observations. The forecast day on the y-axis is the end of the 24-hour period over which the precipitation is accumulated. The threshold is chosen to reflect the forecast skill that is achieved at approximately day 5.5 at the beginning of the strategy period.

# Supplementary headline measure (i): 1 - SEEPS for 24-hr deterministic Precipitation



Supplementary headline score for **deterministic** precipitation forecasts. The curve shows the number of days for which the centered 12-month mean skill remains above a specified threshold for precipitation forecasts over the extra-tropics. The verification is for 24-hour total precipitation verifying against available synoptic observations. The forecast day on the y-axis is the end of the 24-hour period over which the precipitation is accumulated. The threshold is chosen to reflect the forecast skill that is achieved at approximately **day 3.5** at the beginning of the strategy period.

# Supplementary headline measure (i): 1 - SEEPS

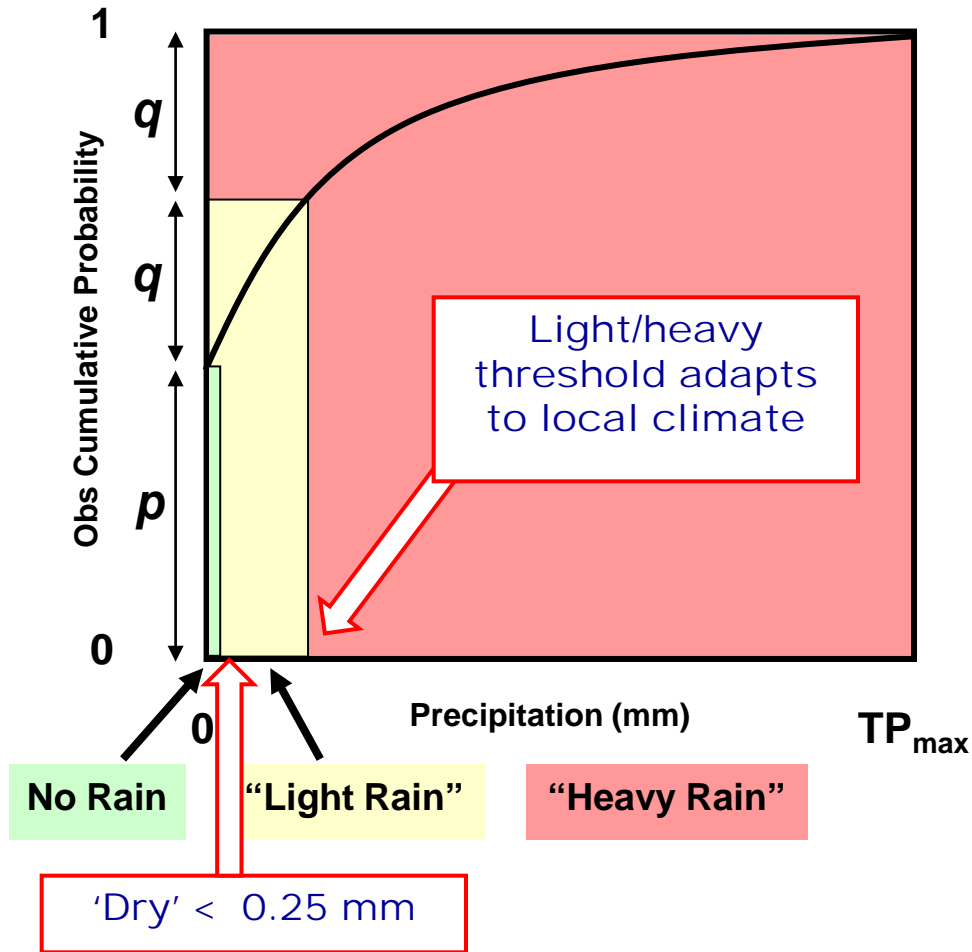
SEEPS  $\Leftrightarrow$  Stable Equitable Error in Probability Space

- ✓ Rodwell et al, 2010: QJRMS, 136 + [Latest ECMWF Newsletter # 128](#)
- ✓ Derived from LEPS score  $\Leftrightarrow$  Linear Error in Probability Space
  - Forecast error is measured in probability space using the climatological cumulative distribution function
- ✓ At each observation location, precipitation is partitioned into 3 categories: (i) “dry” (ii) “light precip” (iii) “heavy precip”
  - Long-term climatological precipitation categories at given SYNOP stations are derived  $\Leftrightarrow$  Accounts for climate differences between stations
- ✓ Evaluates forecast performance across all 3 categories
- ✓ Stable to sample variations and obs error  $\Leftrightarrow$  Good for detecting trends
- ✓ Negatively oriented error measure  $\Leftrightarrow$  Perfect score = 0  $\Leftrightarrow$  1 - SEEPS



# Supplementary headline measure (i): 1 - SEEPS

## Use of Cumulative Distribution



## The characteristics and benefits of SEEPS


A

*Stable:* SEEPS is designed to be as insensitive as possible to sampling uncertainty (for sufficiently skilful forecast systems). This allows more accurate trends to be extracted from noisy data.

*Equitable Error:* A perfect forecast has a SEEPS score of 0. The expected score increases linearly with the unskilled component of the forecast towards a maximum value of 1.

*Probability Space:* This is used to define precipitation categories; SEEPS adapts to the underlying climate to assess the pertinent aspects of local weather. It can be aggregated over heterogeneous climate regions.

# Supplementary headline measure (i): 1 - SEEPS

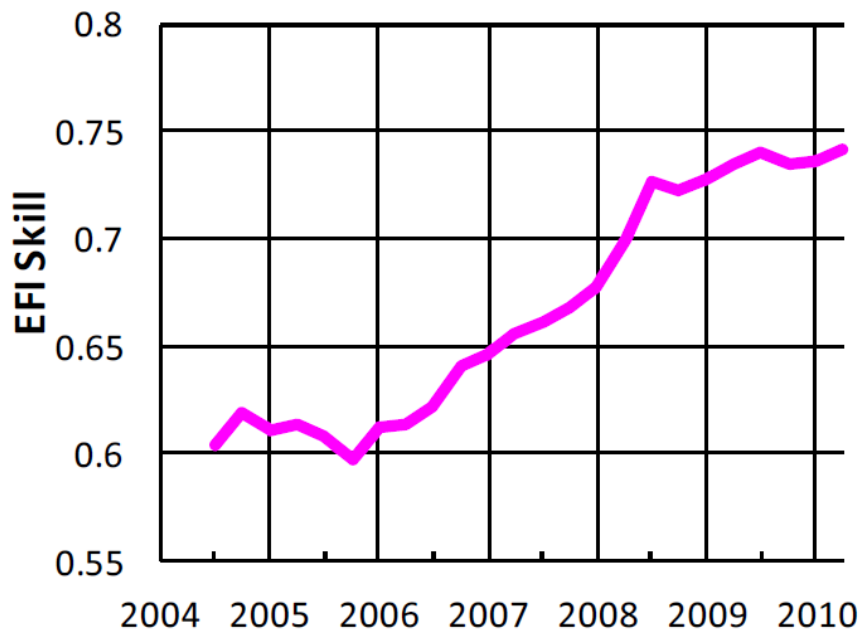
 Year	Predictability, hours	#	
Case FMI Finland	2004	76.6	53629
	2005	77.1	58250
	2006	74.7	39517
	2007	85.1	32160
	2008	86.2	35059
	2009	91.5	56573
	2010	96	64349
	Status 10.9.2011	<b>101.2</b>	58590
	Target 2011	<b>96</b>	?
	Target 2015	<b>106</b>	?

The length in hours for which the mean skill remains above a specified threshold for 24-hour precipitation forecasts in Finland

# Supplementary headline measures (iii) and (iv) for severe weather

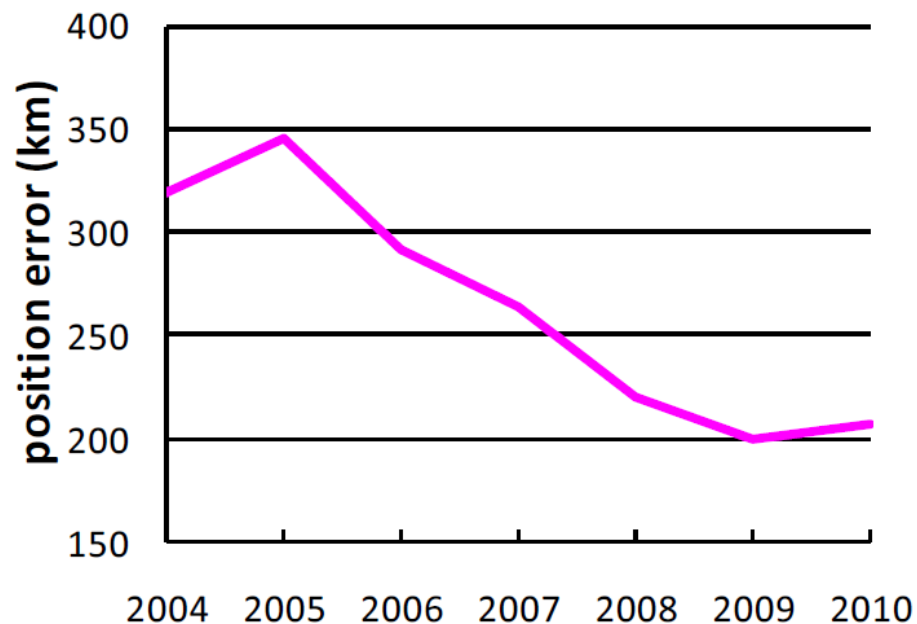
## Proposal

10m wind speed, D+4  
ROC Area



EFI: Extreme event is taken as an observation exceeding 95<sup>th</sup> percentile of station climate; 4-season running mean

Tropical cyclones, D+3  
Mean position error (km)



4-season running mean



New  **ECMWF** Headline  
Verification Measures

Thank You

*Acknowledgement to all 12 Subgroup members and ECMWF staff*