Evaluation of ensemble forecasts

David Richardson

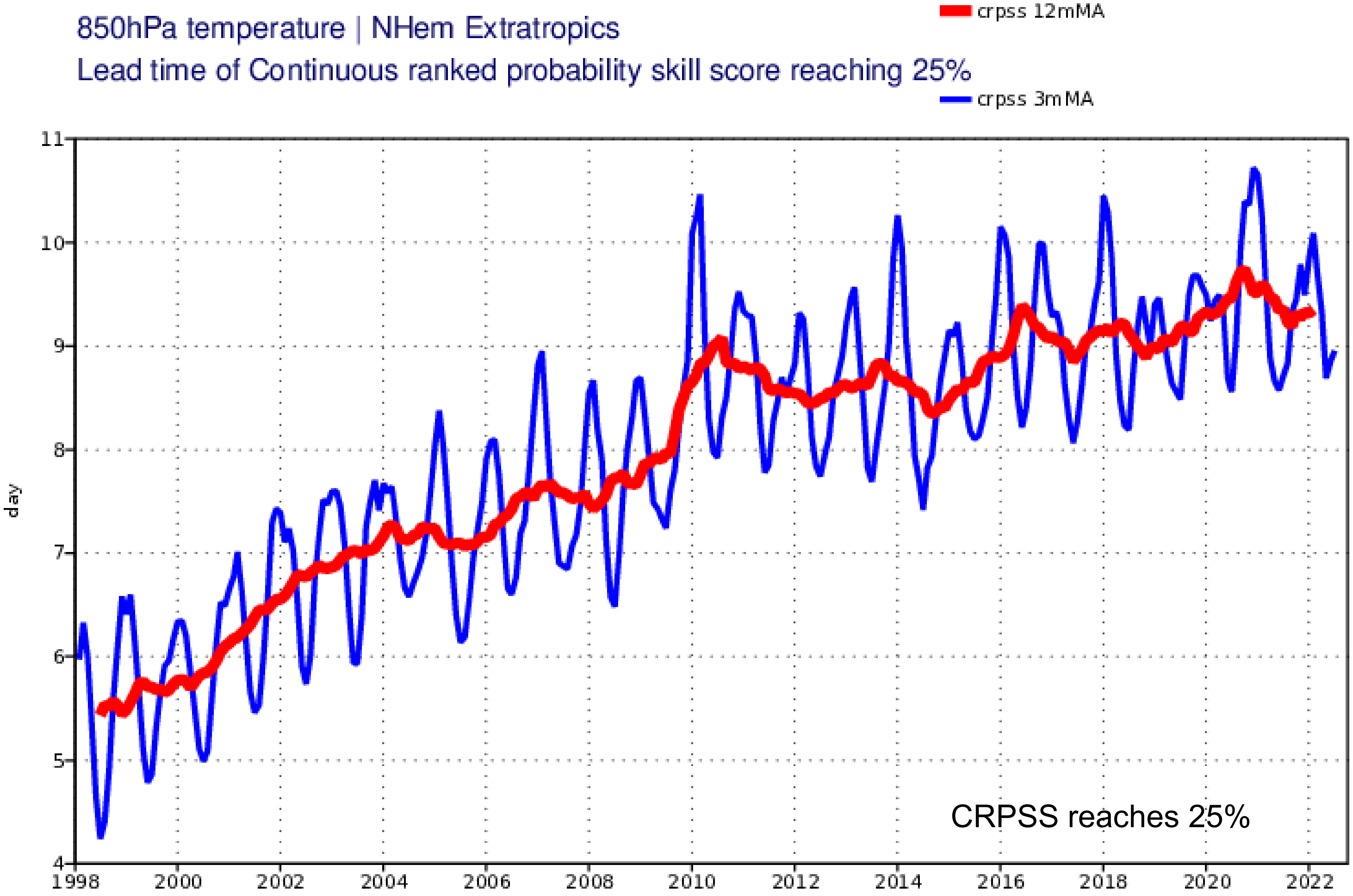
ECMWF

david.richardson@ecmwf.int

Thanks to: Thomas Haiden, Martin Janousek, Zied Ben Bouallegue

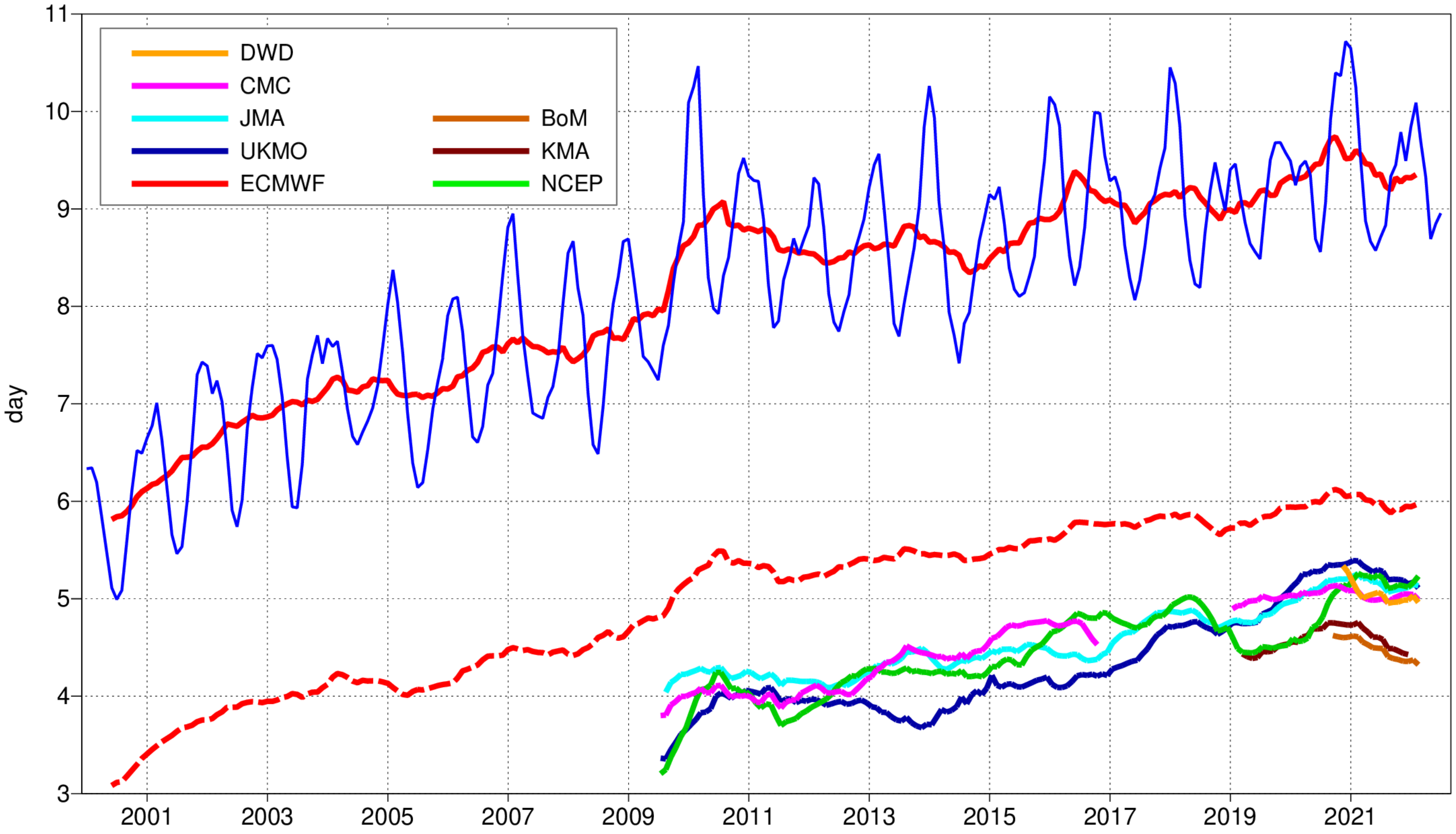
ECMWF

ENS upper-air headline score



ENS upper-air headline score

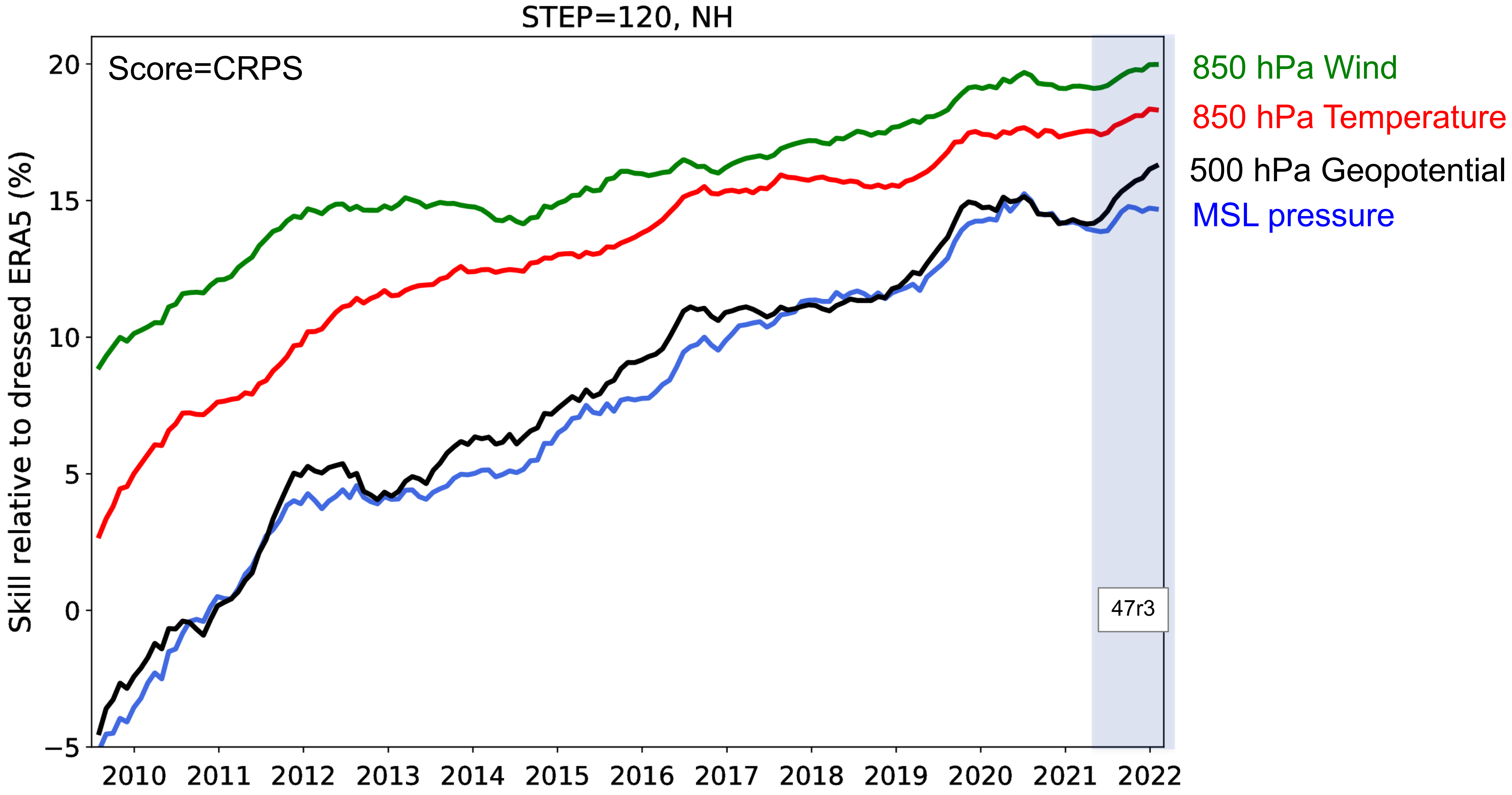
Continuous ranked probability skill score | 850hPa temperature | NHem Extratropics | oper_an enfo beginning



CRPSS reaches 25%

CRPSS reaches 50%

ENS skill relative to dressed ERA5 – Day 5



Decision making - the cost-loss model

- Simplest possible case - but shows many important features
- There are only two important weather types: weather is either “good” or “bad”
- A particular user or decision maker will be affected by bad weather - they have a choice of two actions
  - If they do nothing and bad weather occurs they suffer a loss L
  - However, they can decide to take some protective action to prevent this possible loss, but it will cost C

	event occurs	event does NOT occur
Protection: YES	**C**	**C**
Protection: NO	**L**	**0**

- If no forecast just use climatological information
  - Always protect (if often occurs)
  - Never protect (if rarely occurs)

Value of deterministic forecasts

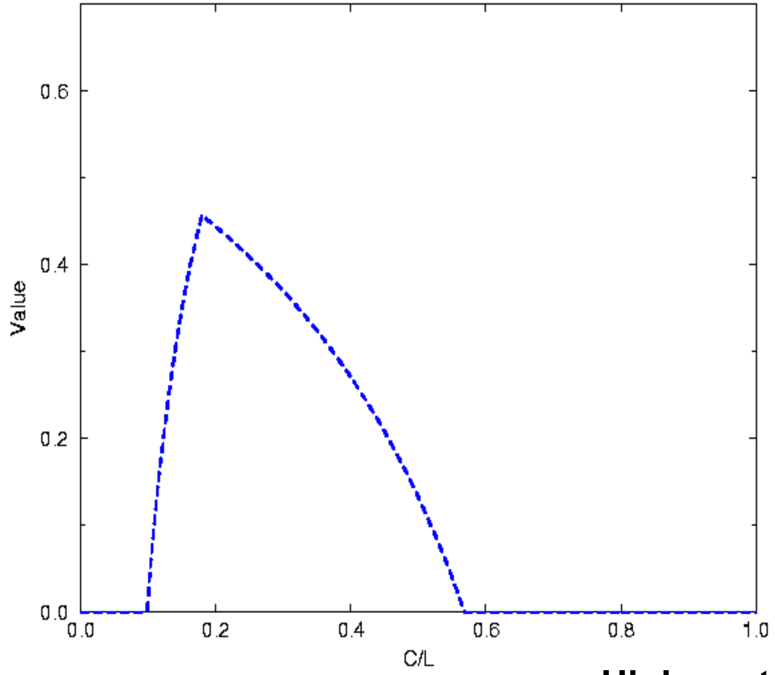
- Using forecast: protect when event is forecast
- Value

$$V = \frac{\text{saving from using forecast}}{\text{saving from perfect forecast}}$$

- V = 0 forecast is no better than climate
- V = 1 forecast is perfect (no misses, no false alarms)

- Value depends on forecast quality: Hits and False alarms
  - but value also depends on the user (C/L)
  - and on the weather event (base rate, ō)

	event occurs	event does NOT occur
Protection: YES	**C**	**C**
Protection: NO	**L**	**0**

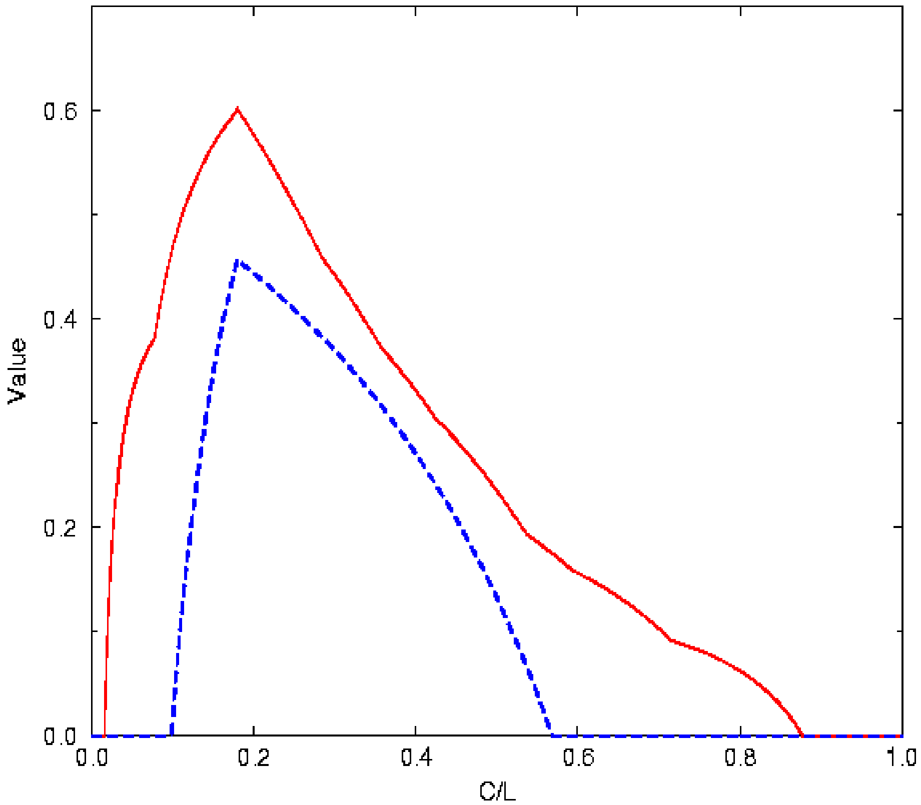


High loss from missed event (hit rate important)

High cost to protect (false alarm rate important)

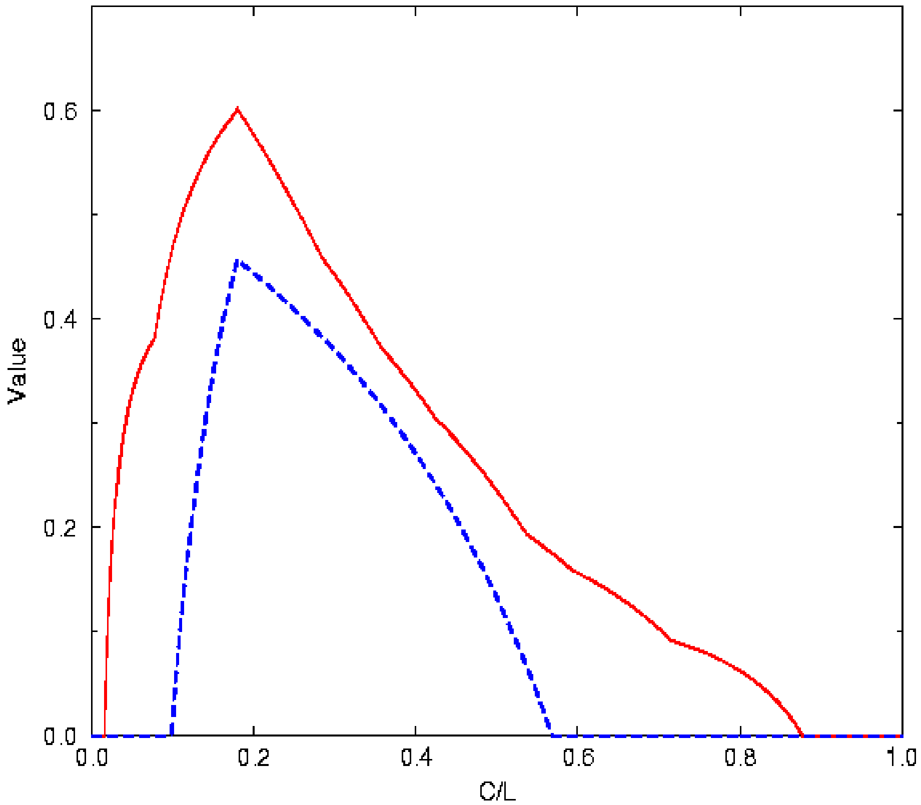
Benefit of probability forecasts

- If the cost of protection is high wait until event is more certain
  - False alarms are more important
- If the loss is greater then protect even at low probability
  - Missed events are more important
- Changing the probability threshold at which to take action gives different hit rates and false alarm rates
- The optimal probability threshold depends on the user: p_t=C/L
- Using the probabilities allows decision makers to take decisive action according to their own risks – these are different for each user
- Even if the user does not have an explicit cost/loss they are still aware of the relative importance of false alarms and missed events



Benefit of probability forecasts

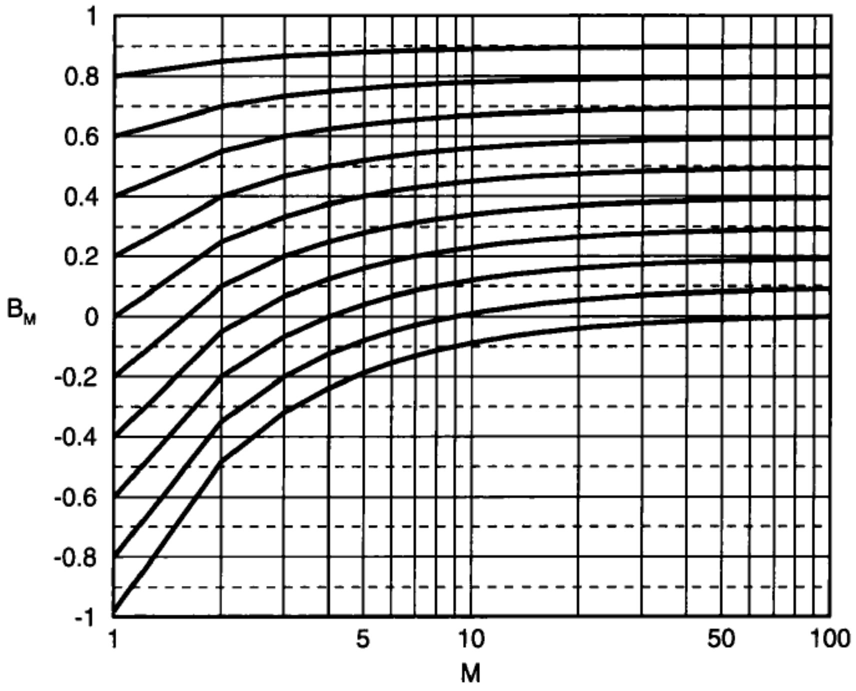
- If the cost of protection is high wait until event is more certain
  - False alarms are more important
- If the loss is greater then protect even at low probability
  - Missed events are more important
- Changing the probability threshold at which to take action gives different hit rates and false alarm rates
- The optimal probability threshold depends on the user: p_t=C/L
- Using the probabilities allows decision makers to take decisive action according to their own risks – these are different for each user
- Even if the user does not have an explicit cost/loss they are still aware of the relative importance of false alarms and missed events

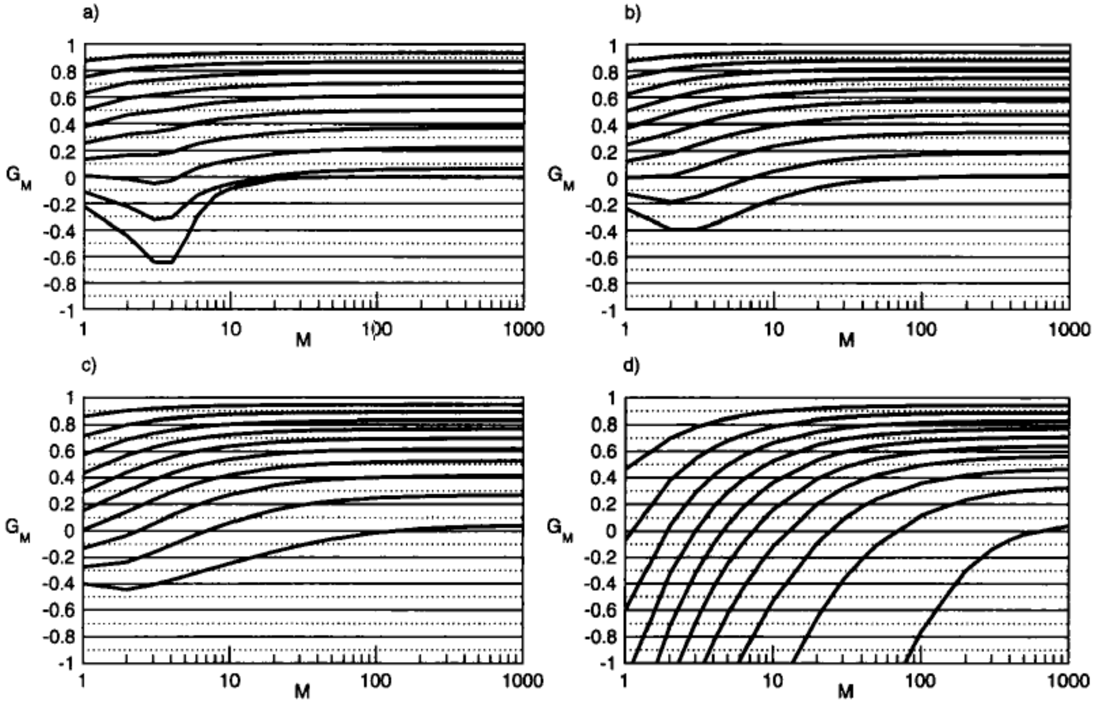


Brier skill score - average value over all C/L

Skill, value and ensemble size

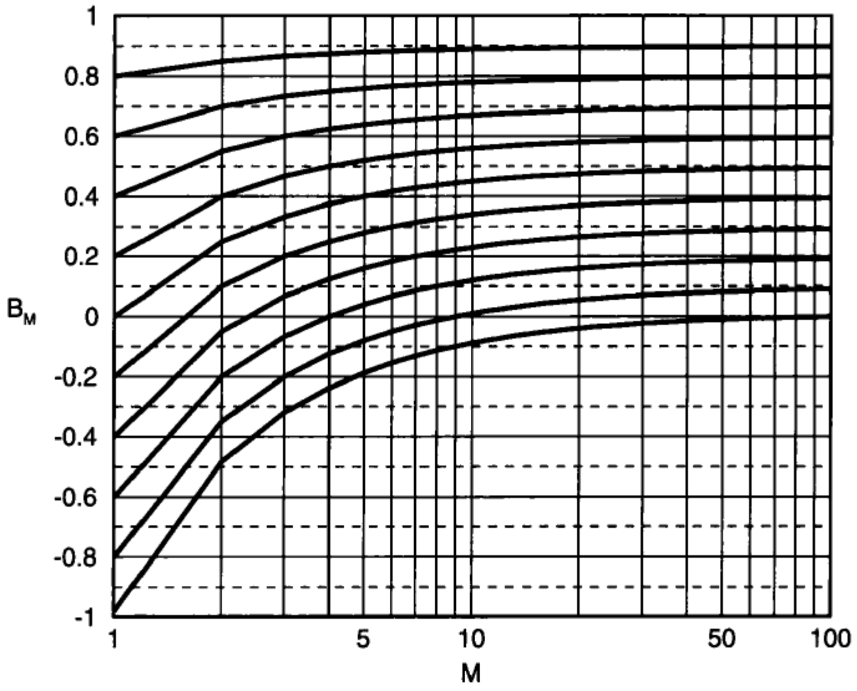
- BS – value for uniform distribution of users (C/L)
- Brier score saturates with ensemble size < 50
  - Assumes underlying pdf is reliable (ie no model biases)
- Is this relevant for users?
- Are larger ensembles any use?
- What about other distributions of users?

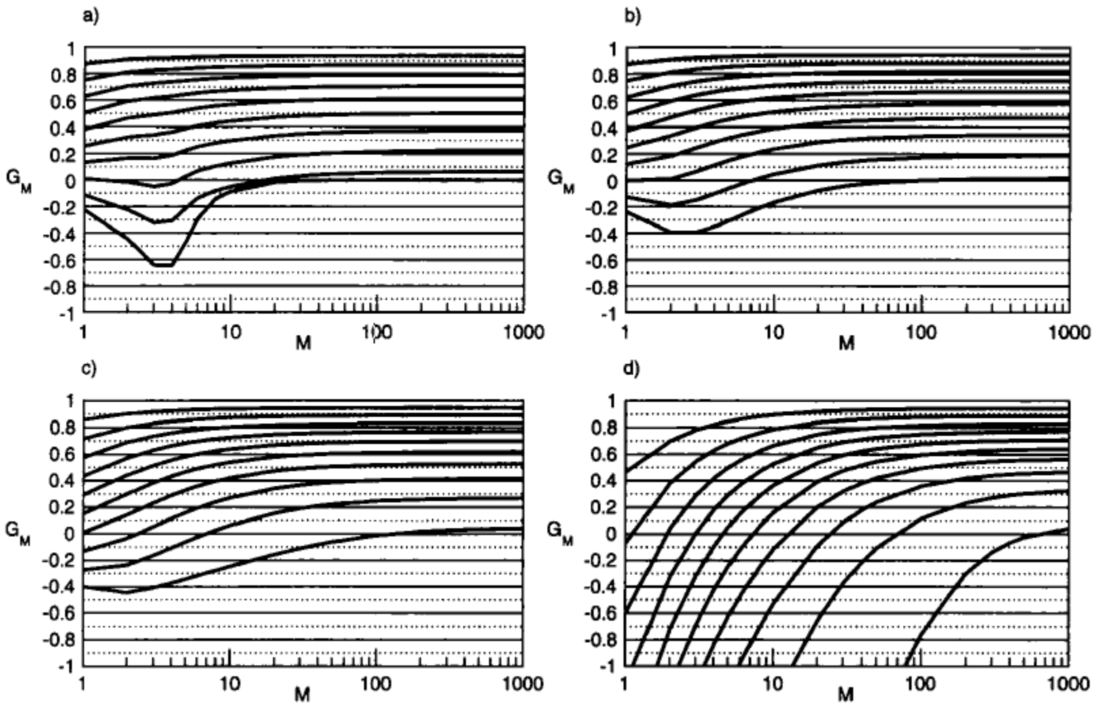


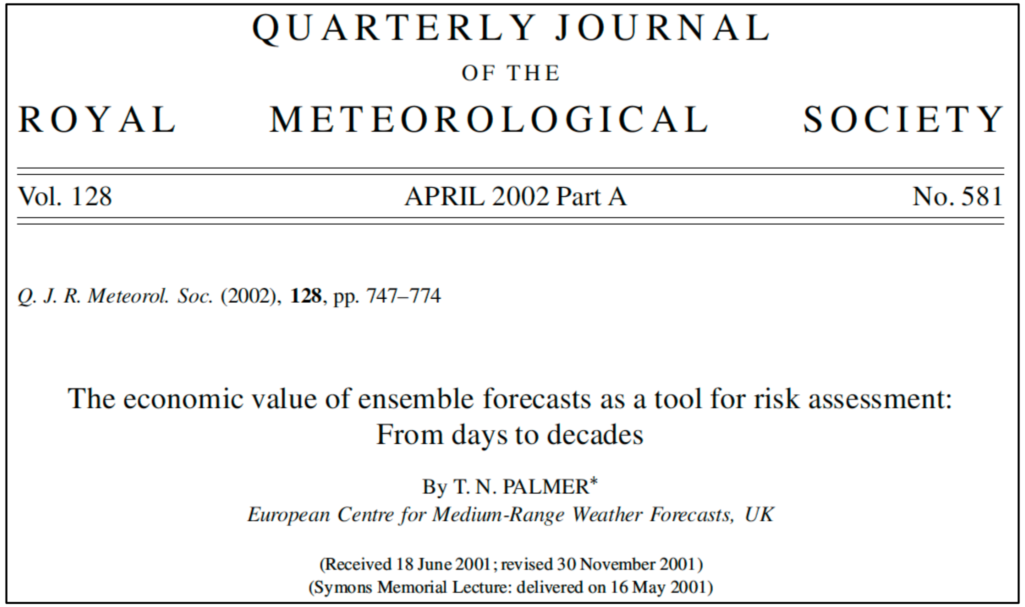


Skill, value and ensemble size

- BS – value for uniform distribution of users (C/L)
- Brier score saturates with ensemble size < 50
  - Assumes underlying pdf is reliable (ie no model biases)
- Is this relevant for users?
- Are larger ensembles any use?
- What about other distributions of users?







QUARTERLY JOURNAL

OF THE

ROYAL METEOROLOGICAL SOCIETY

Vol. 128 APRIL 2002 Part A No. 581

Q. J. R. Meteorol. Soc. (2002), **128**, pp. 747–774

The economic value of ensemble forecasts as a tool for risk assessment:
From days to decades

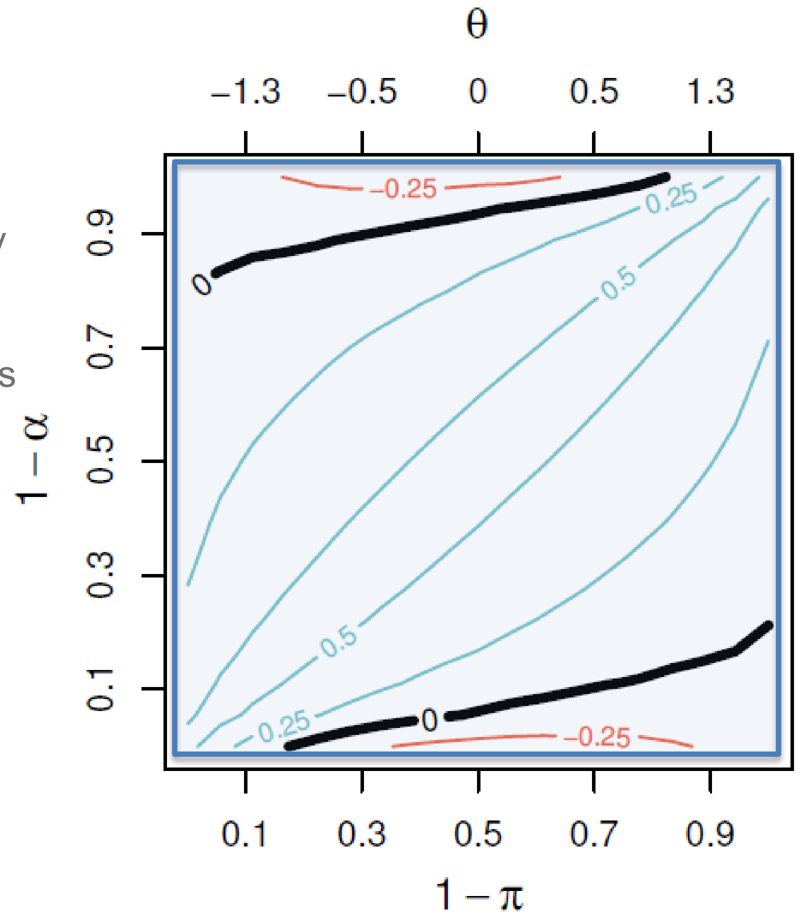
By T. N. PALMER*
European Centre for Medium-Range Weather Forecasts, UK

(Received 18 June 2001; revised 30 November 2001)
(Symons Memorial Lecture: delivered on 16 May 2001)

Value and skill - Summary scores

- **Continuous Ranked probability (skill) score**
  All events and probability levels (cost-loss ratios)

Event threshold



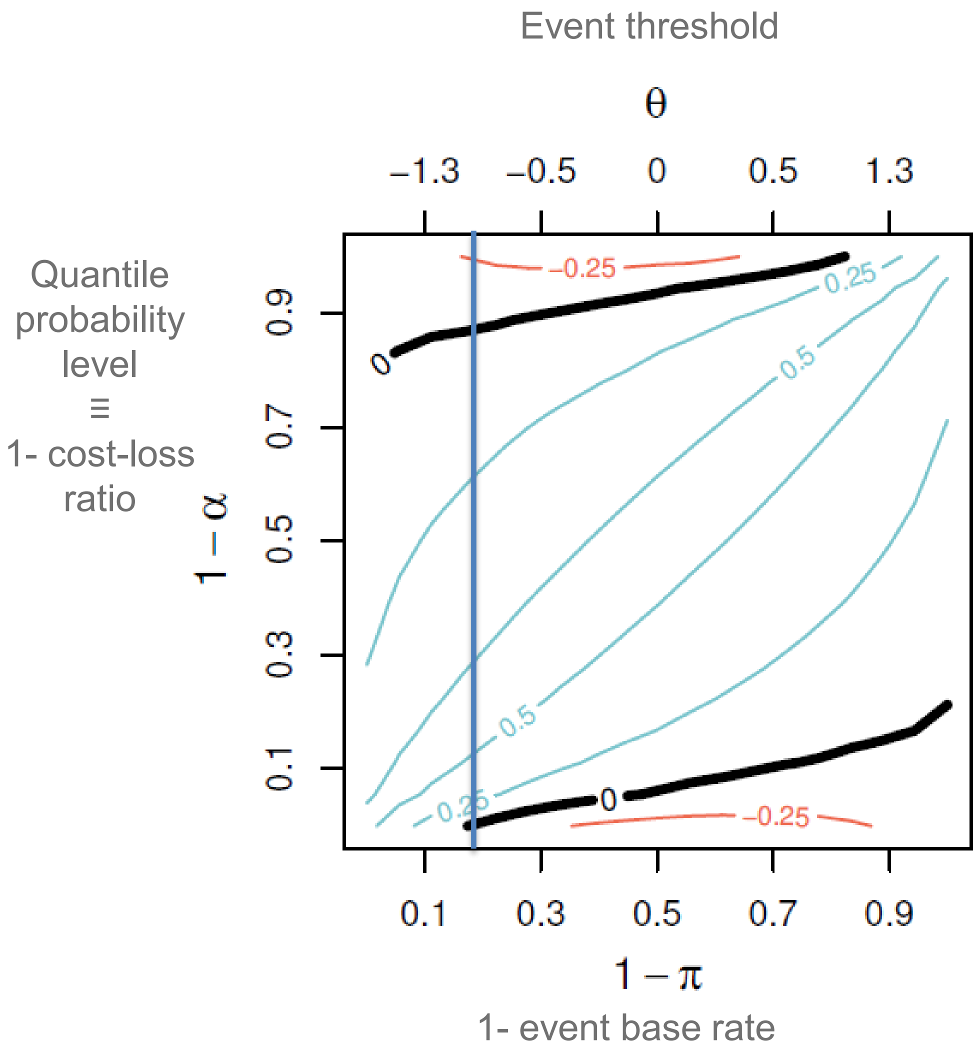
Quantile probability level
≡
1- cost-loss ratio

1- event base rate

Value and skill - Summary scores

- **Continuous Ranked probability (skill) score**
  All events and probability levels (cost-loss ratios)

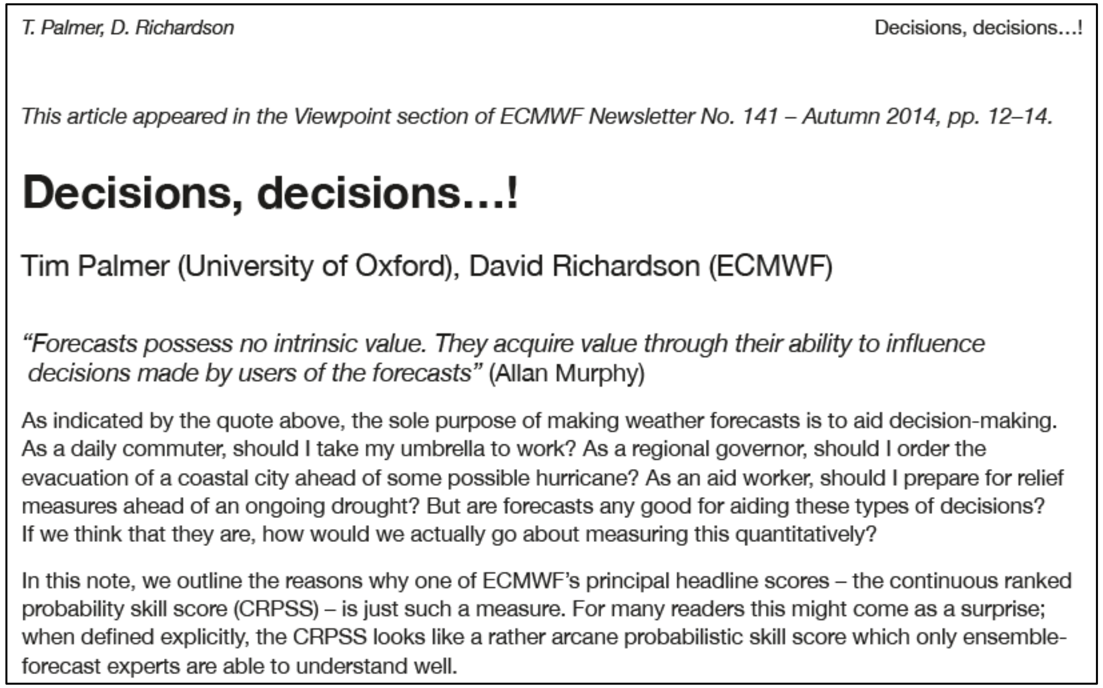
- **Brier (skill) score**
  Fixed event
  Focus on a vertical line



Value and skill - Summary scores

- **Continuous Ranked probability (skill) score**
  All events and probability levels (cost-loss ratios)

- **Brier (skill) score**
  Fixed event
  Focus on a vertical line



T. Palmer, D. Richardson — Decisions, decisions…!

This article appeared in the Viewpoint section of ECMWF Newsletter No. 141 – Autumn 2014, pp. 12–14.

Decisions, decisions…!

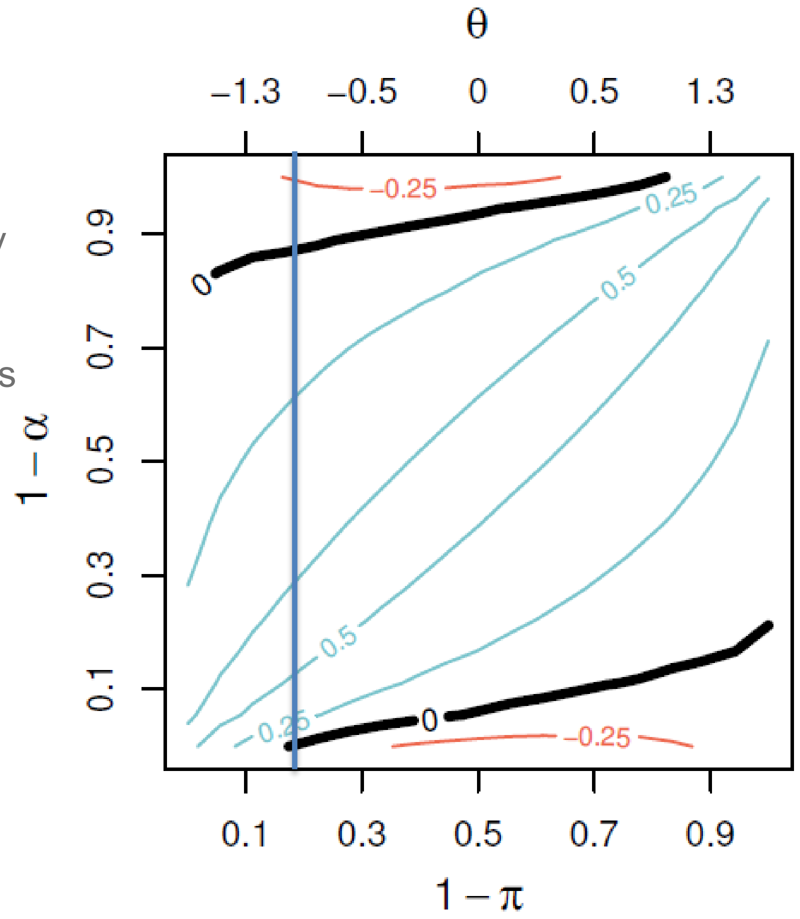
Tim Palmer (University of Oxford), David Richardson (ECMWF)

"Forecasts possess no intrinsic value. They acquire value through their ability to influence decisions made by users of the forecasts" (Allan Murphy)

As indicated by the quote above, the sole purpose of making weather forecasts is to aid decision-making. As a daily commuter, should I take my umbrella to work? As a regional governor, should I order the evacuation of a coastal city ahead of some possible hurricane? As an aid worker, should I prepare for relief measures ahead of an ongoing drought? But are forecasts any good for aiding these types of decisions? If we think that they are, how would we actually go about measuring this quantitatively?

In this note, we outline the reasons why one of ECMWF's principal headline scores – the continuous ranked probability skill score (CRPSS) – is just such a measure. For many readers this might come as a surprise; when defined explicitly, the CRPSS looks like a rather arcane probabilistic skill score which only ensemble-forecast experts are able to understand well.

Event threshold



Quantile probability level ≡ 1- cost-loss ratio

1- event base rate

Value and skill - Summary scores

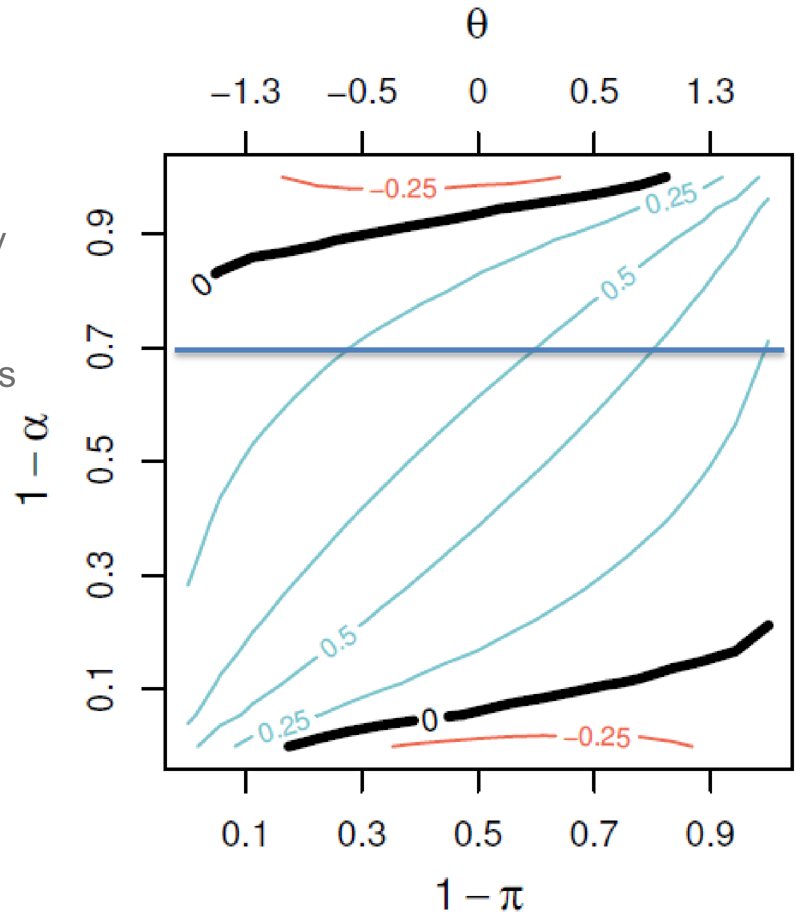
- **Continuous Ranked probability (skill) score**
  All events and probability levels (cost-loss ratios)

- **Brier (skill) score**
  Fixed event
  Focus on a vertical line

- **Quantile (skill) score**
  Fixed probability level
  Focus on a horizontal line

Event threshold

Quantile probability level ≡ 1- cost-loss ratio



1- event base rate

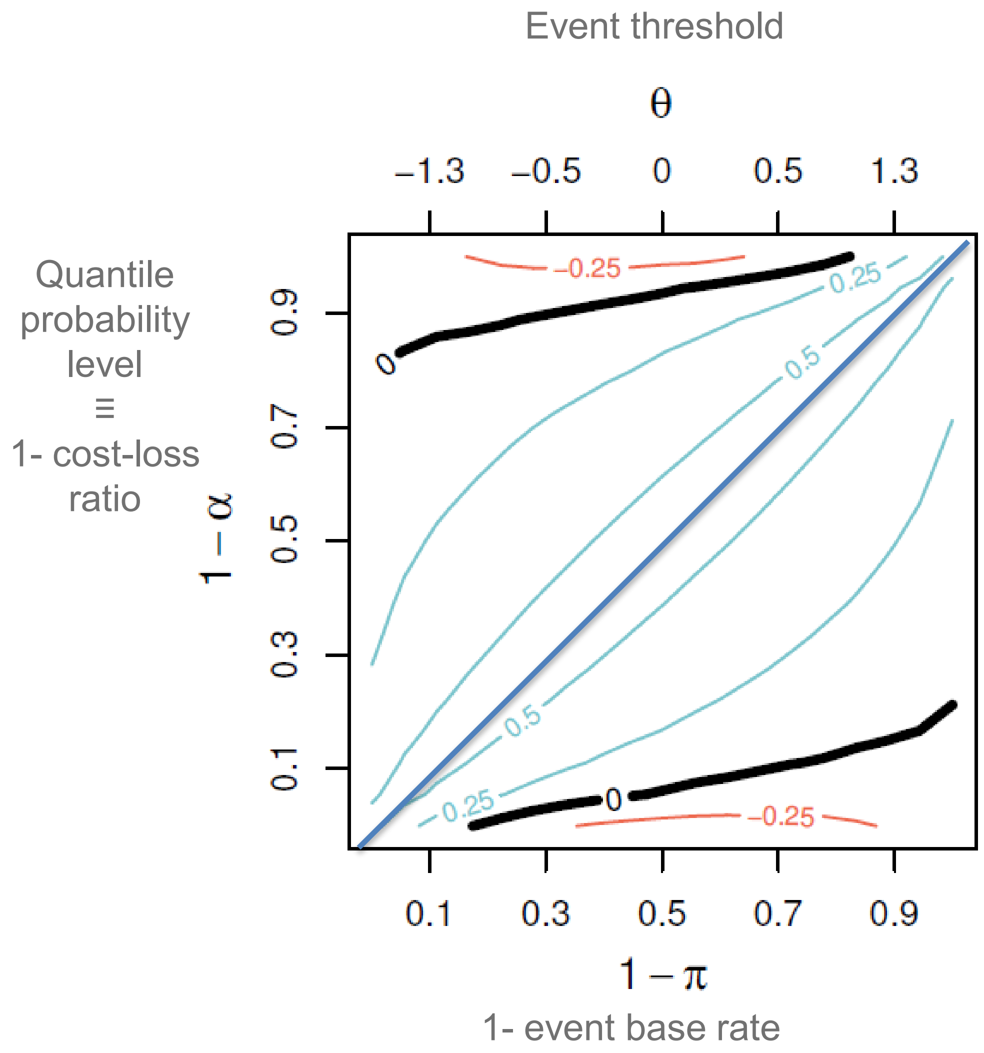
Value and skill - Summary scores

- **Continuous Ranked probability (skill) score**
  All events and probability levels (cost-loss ratios)

- **Brier (skill) score**
  Fixed event
  Focus on a vertical line

- **Quantile (skill) score**
  Fixed probability level
  Focus on a horizontal line

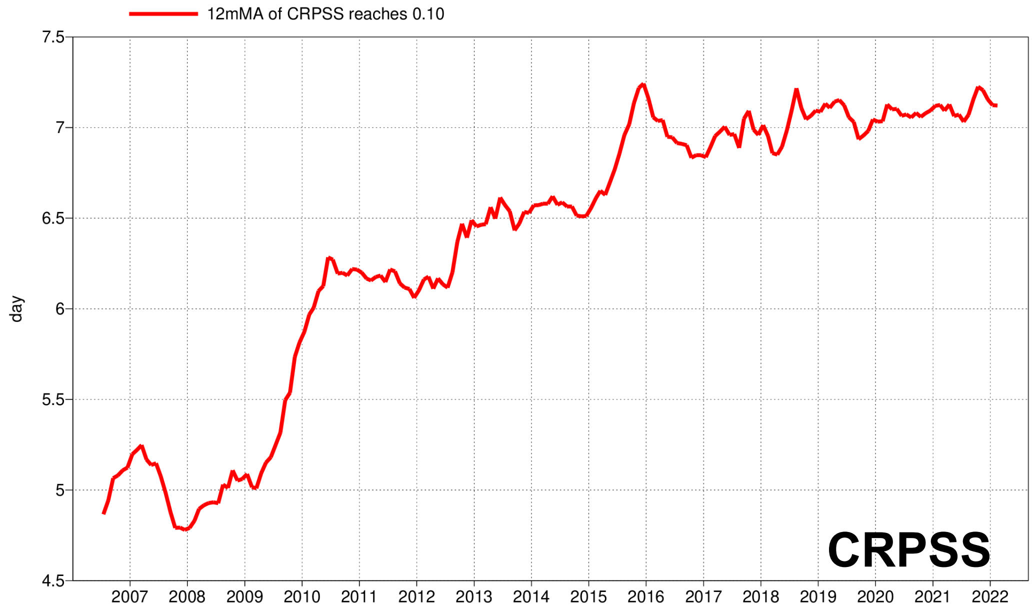
- **Diagonal (skill) score**
  Event base rate and probability level directly related
  Focus on the ascendant diagonal

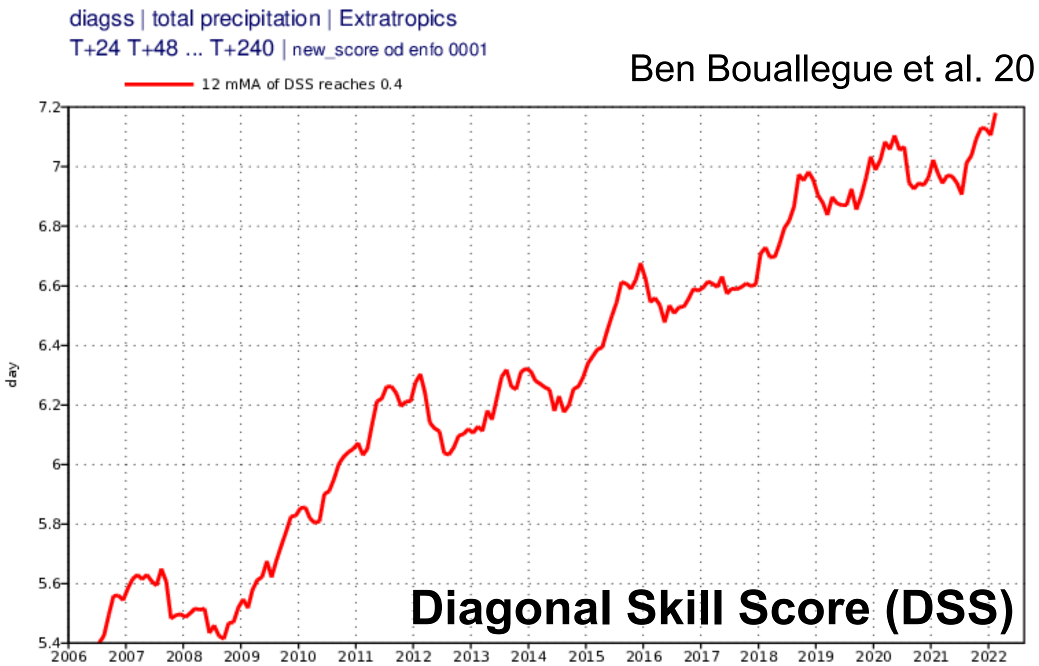


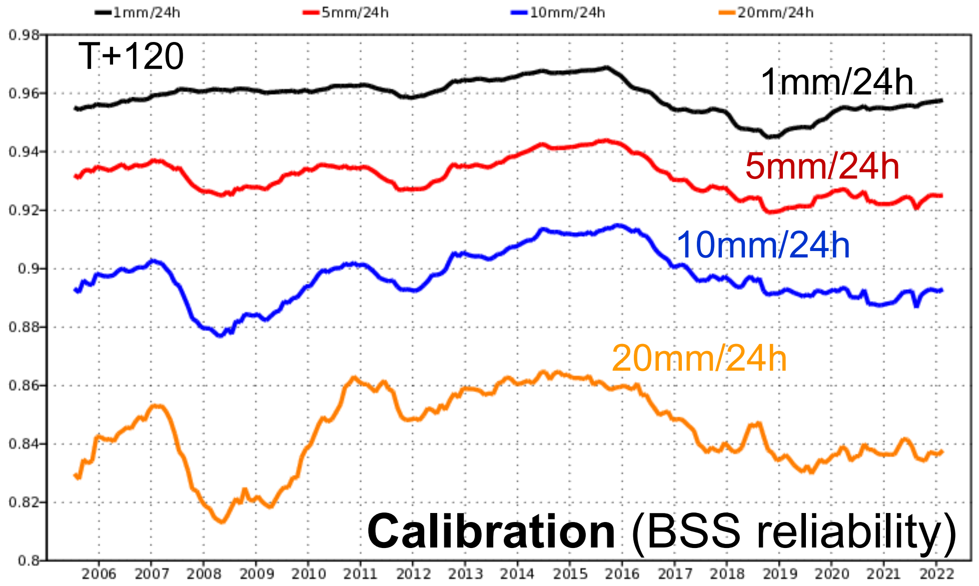
“The diagonal score: definition, properties, and interpretations”, Ben Bouallegue et al. 2018 QJRMS

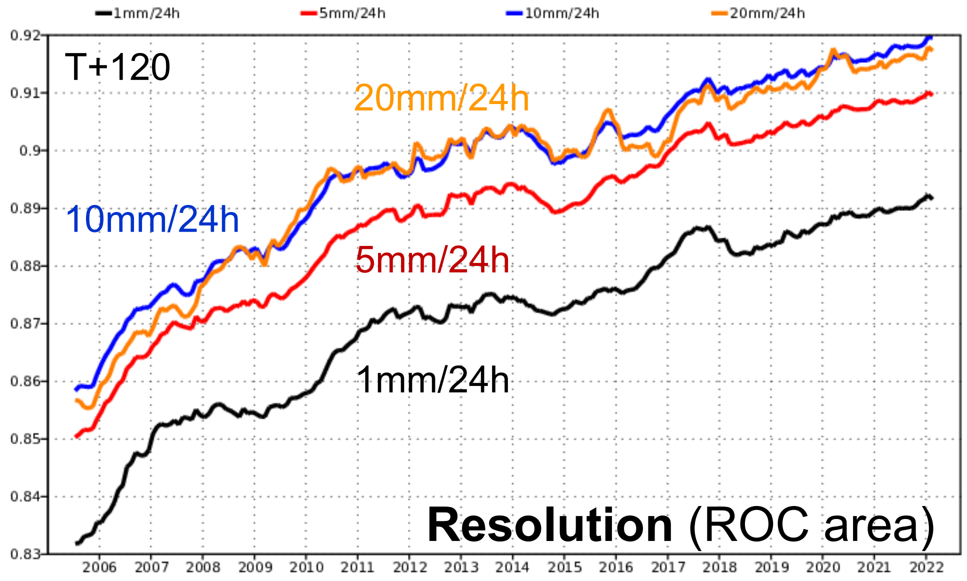
ENS precipitation scores

Ben Bouallegue et al. 2018 QJRMS





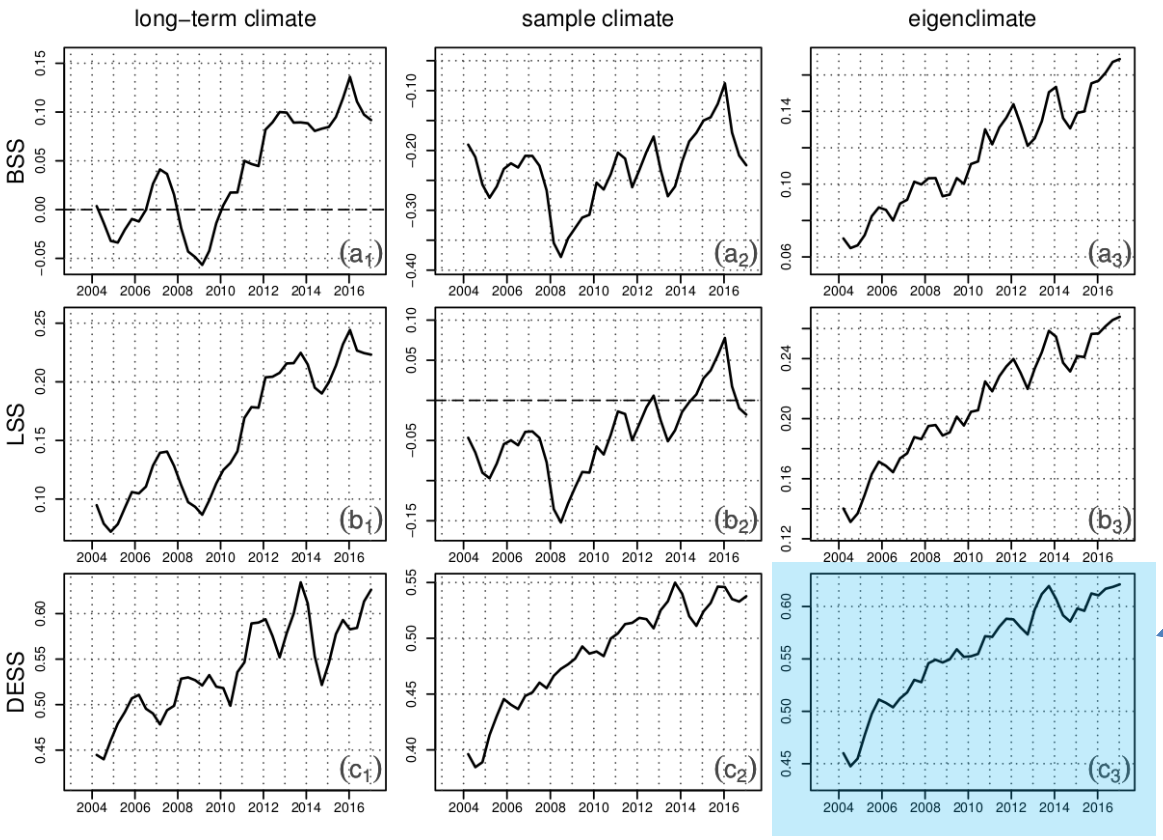




High-impact weather: trends in ensemble forecast performance

Comparison of performance evolution using

- **different scores** (Brier score, Logarithmic score, Diagonal elementary score)
- **different climatology** definitions



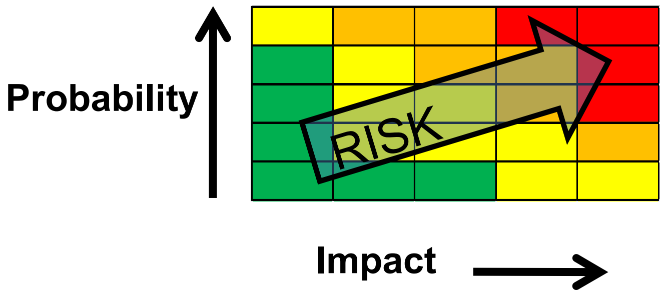
Example:
24 h precipitation
95th percentile
Day 5

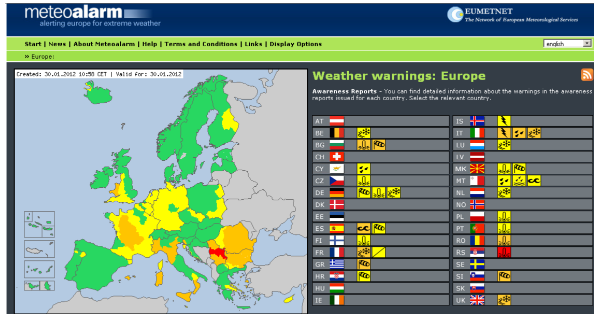
Highest signal-to-noise ratio

“Monitoring trends in ensemble forecast performance focusing on surface variables and high-impact events”, Ben Bouallegue et al. 2019 QJRMS

Summary – skill and value of ensemble forecasts

- Ensembles forecasts provide an explicit, detailed representation of model uncertainties, and potential of unusual events
- Supporting decision making: societal and economic value of forecasts
- Forecasts only have value if people use them
  - make a decision or take an action which would not otherwise have been made
- To make a good decision need to know the probability and the impact (consequence to the individual user)
- Value of forecasts depends on both the quality of the forecasting system and the costs/losses of the user
- Skill scores are summary measures, sometimes with implicit assumptions about distributions of users




meteoalarm
alerting europe for extreme weather

EUMETNET
The Network of European Meteorological Services

Start | News | About Meteoalarm | Help | Terms and Conditions | Links | Display Options

english

» Europe:

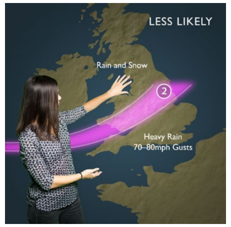
Created: 30.01.2012 10:58 CET | Valid for: 30.01.2012

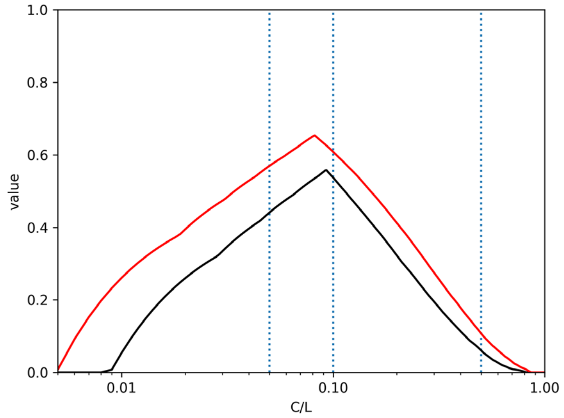
Weather warnings: Europe

Awareness Reports - You can find detailed information about the warnings in the awareness reports issued for each country. Select the relevant country.

AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GR, HR, HU, IE

IS, IT, LU, LV, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, UK

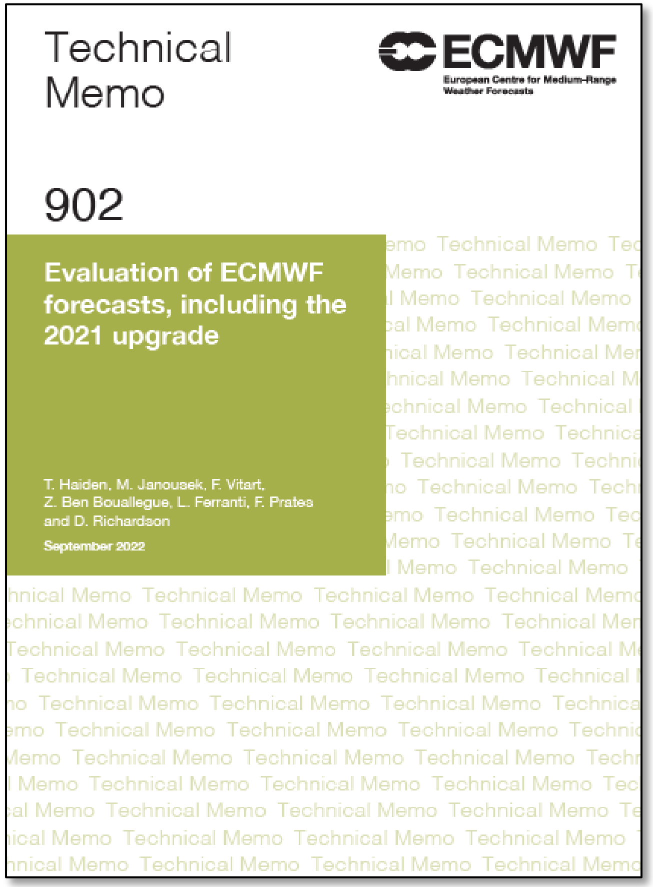




Forecast performance

- 8 headline scores
  - HRES and ENS upper-air skill
  - HRES and ENS precipitation
  - Severe weather: TC position and EFI for extreme wind
  - Frequency of large temperature errors
  - Weekly mean 2m temperature (terciles)
- Comparison with reference systems
- Comparison with other centres
- Evaluation for severe weather
- Additional verification and in-depth diagnostics
- See ECMWF web site for latest results

www.ecmwf.int/en/forecasts/quality-our-forecasts



Technical Memo

902

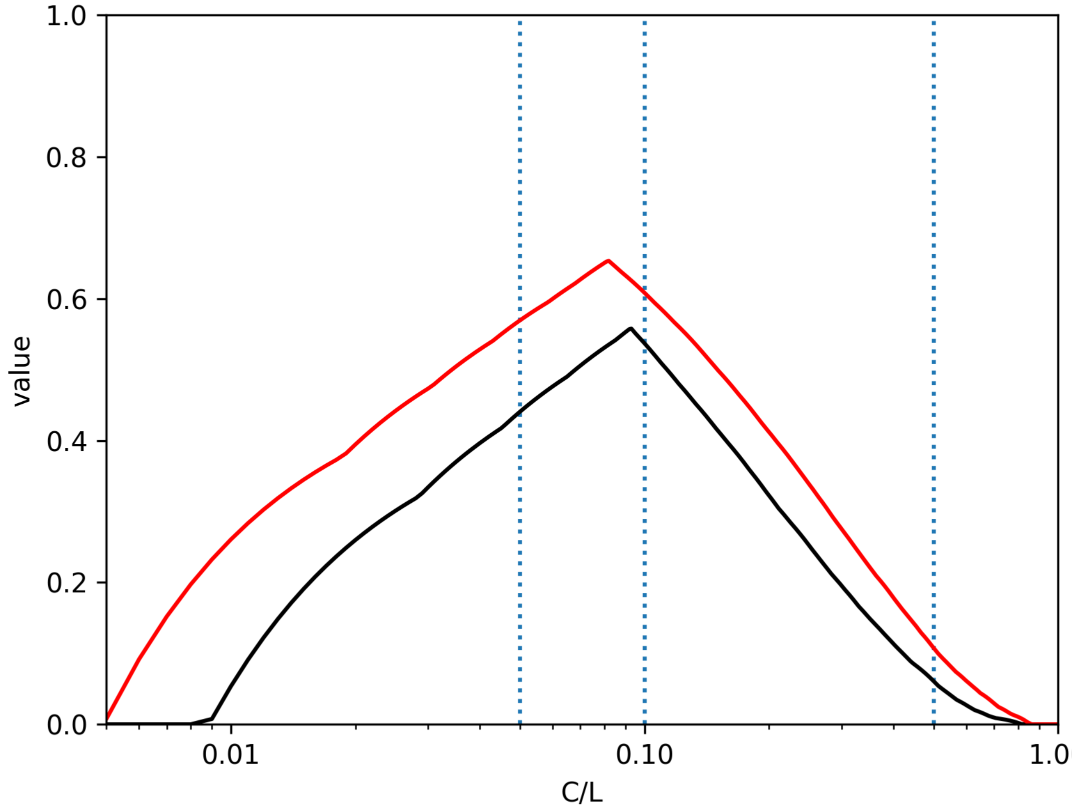
Evaluation of ECMWF forecasts, including the 2021 upgrade

T. Haiden, M. Janousek, F. Vitart, Z. Ben Bouallegue, L. Ferranti, F. Prates and D. Richardson

September 2022

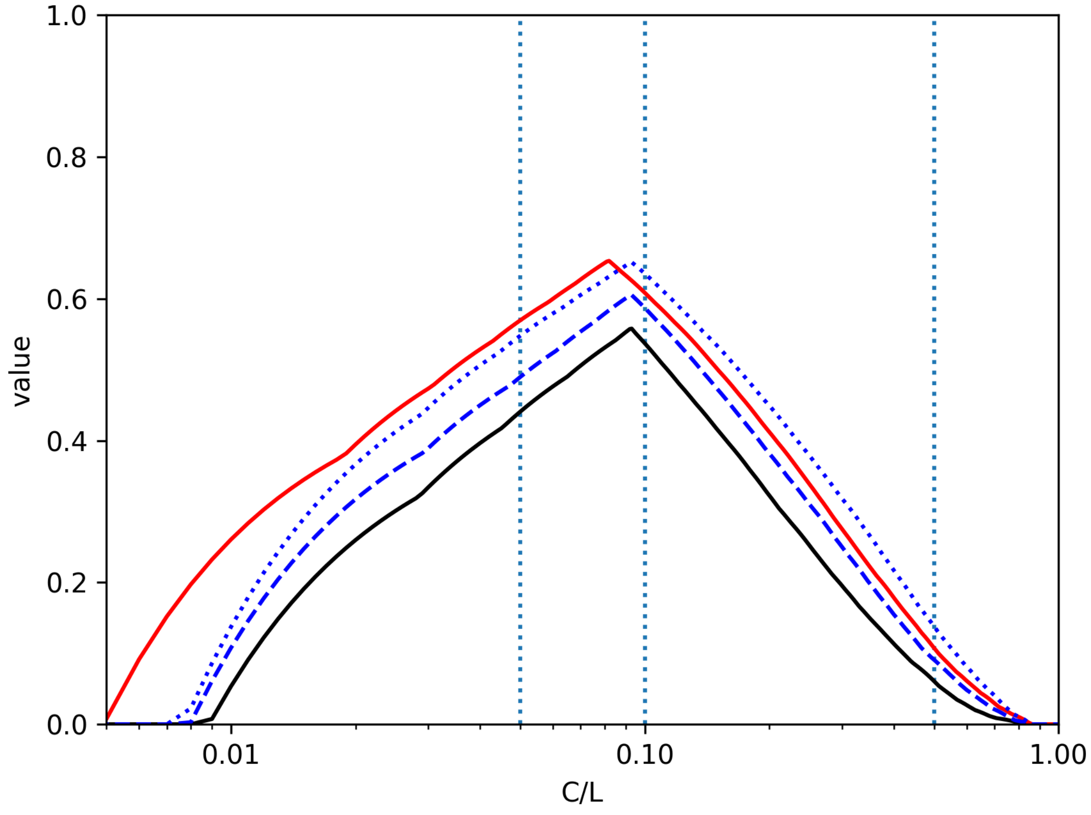
Value of precipitation forecasts

- 7-day forecasts (n hemisphere) 24h precipitation > 5mm
- 2022 (red) v 2010 (black)



Value of precipitation forecasts

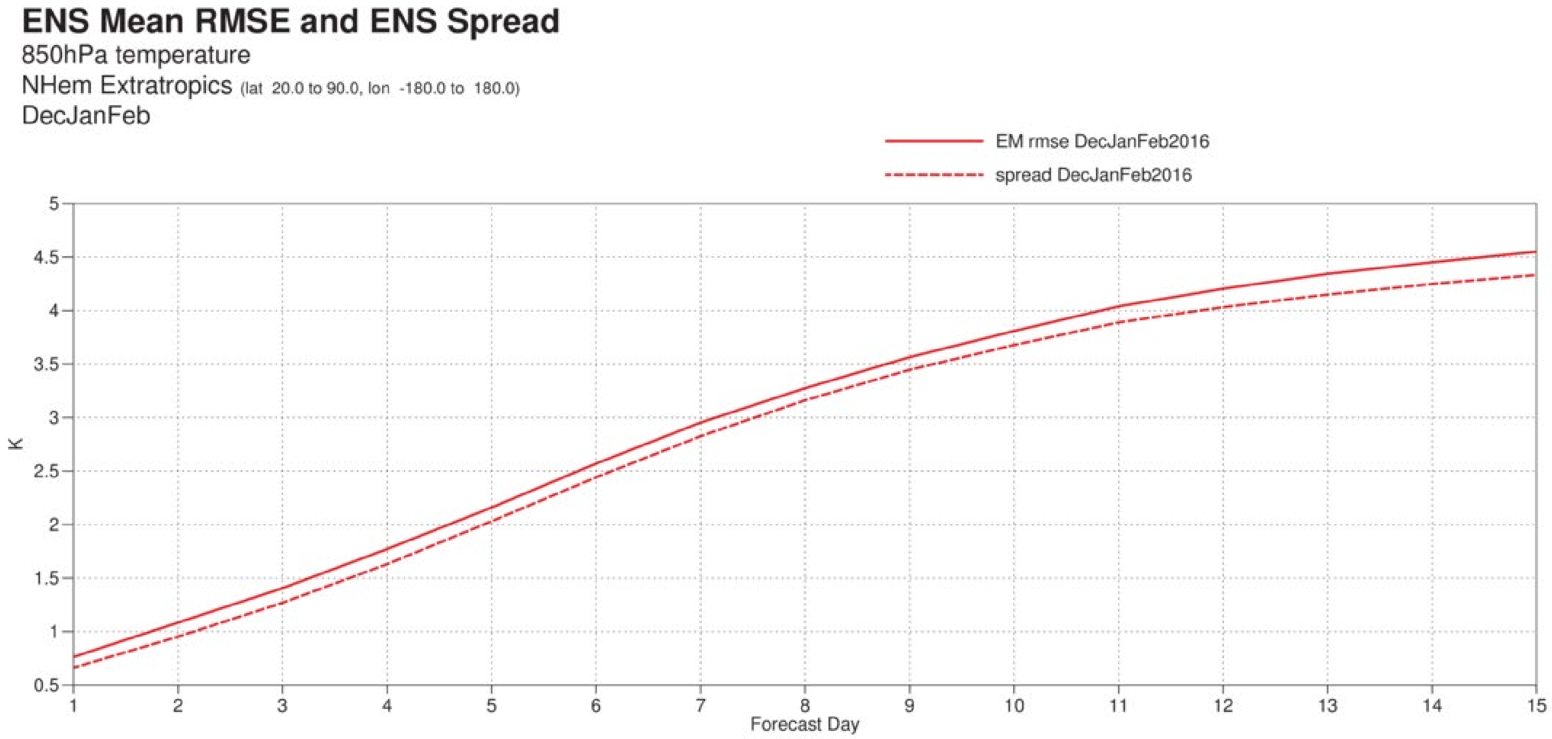
- 7-day forecasts (n hemisphere) 24h precipitation > 5mm
- 2022 (red) v 2010 (black)

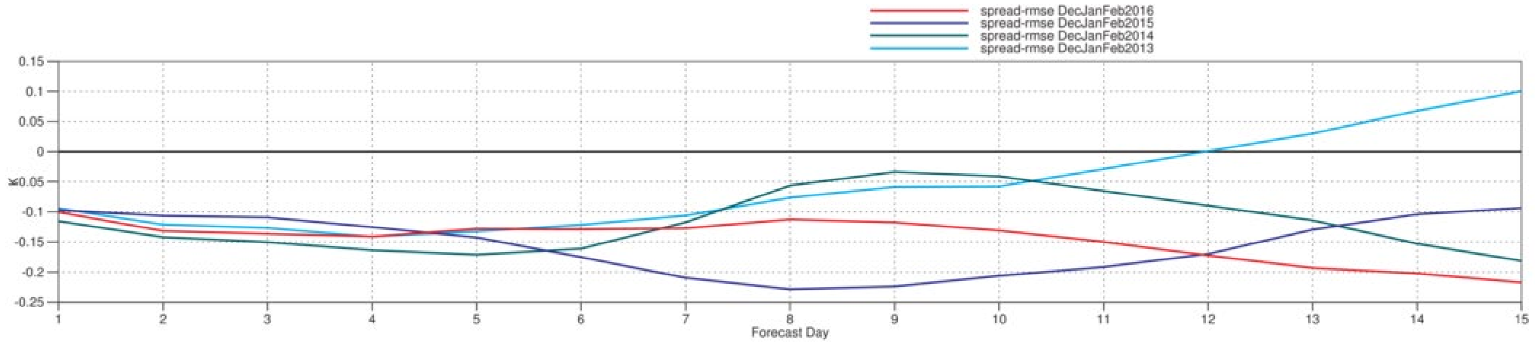


ENS spread and error

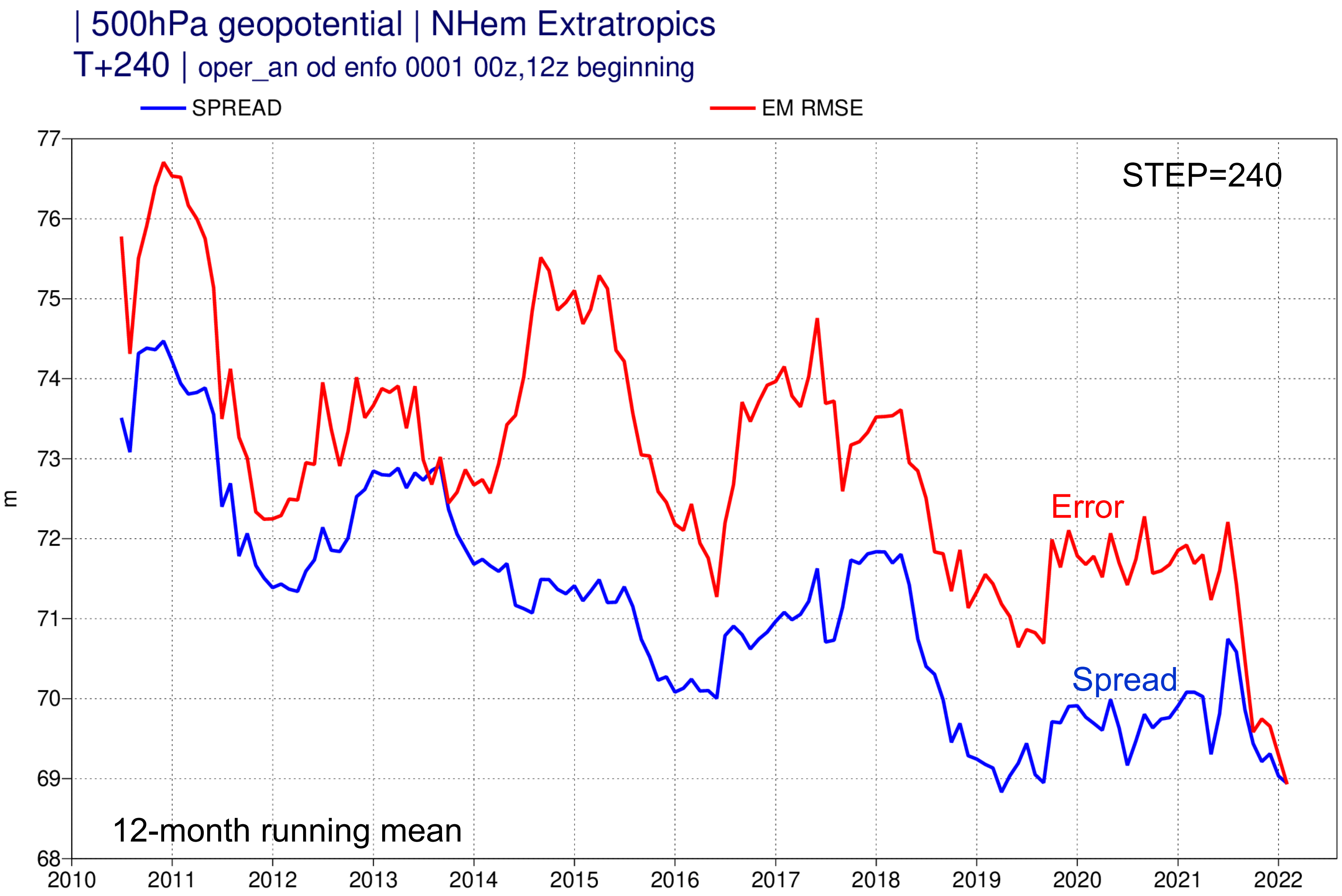
850 hPa temperature, Northern Hemisphere

ENS spread (dashed),
RMS error of ensemble-mean (full lines),
and their difference (below) in winter.

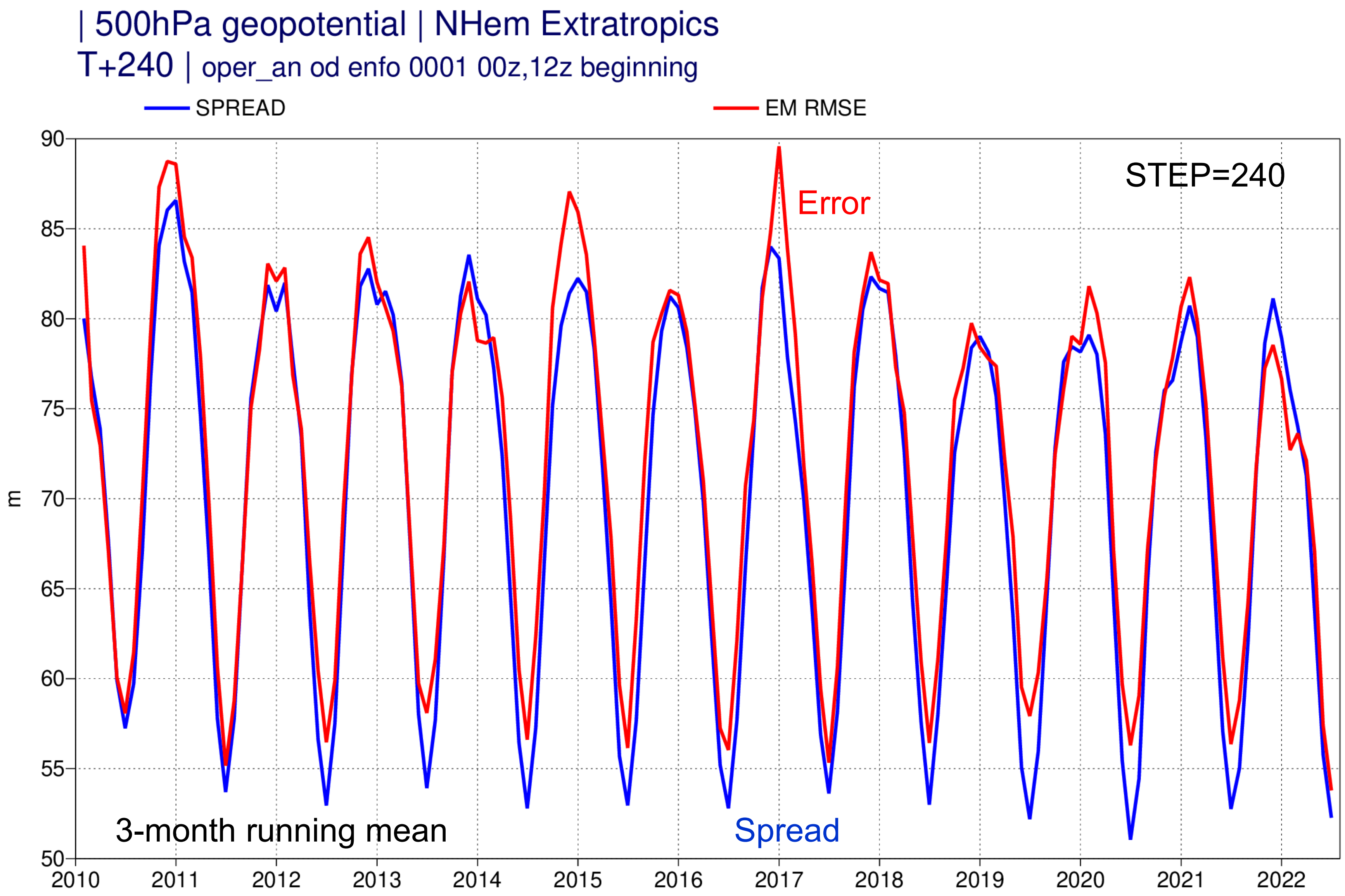




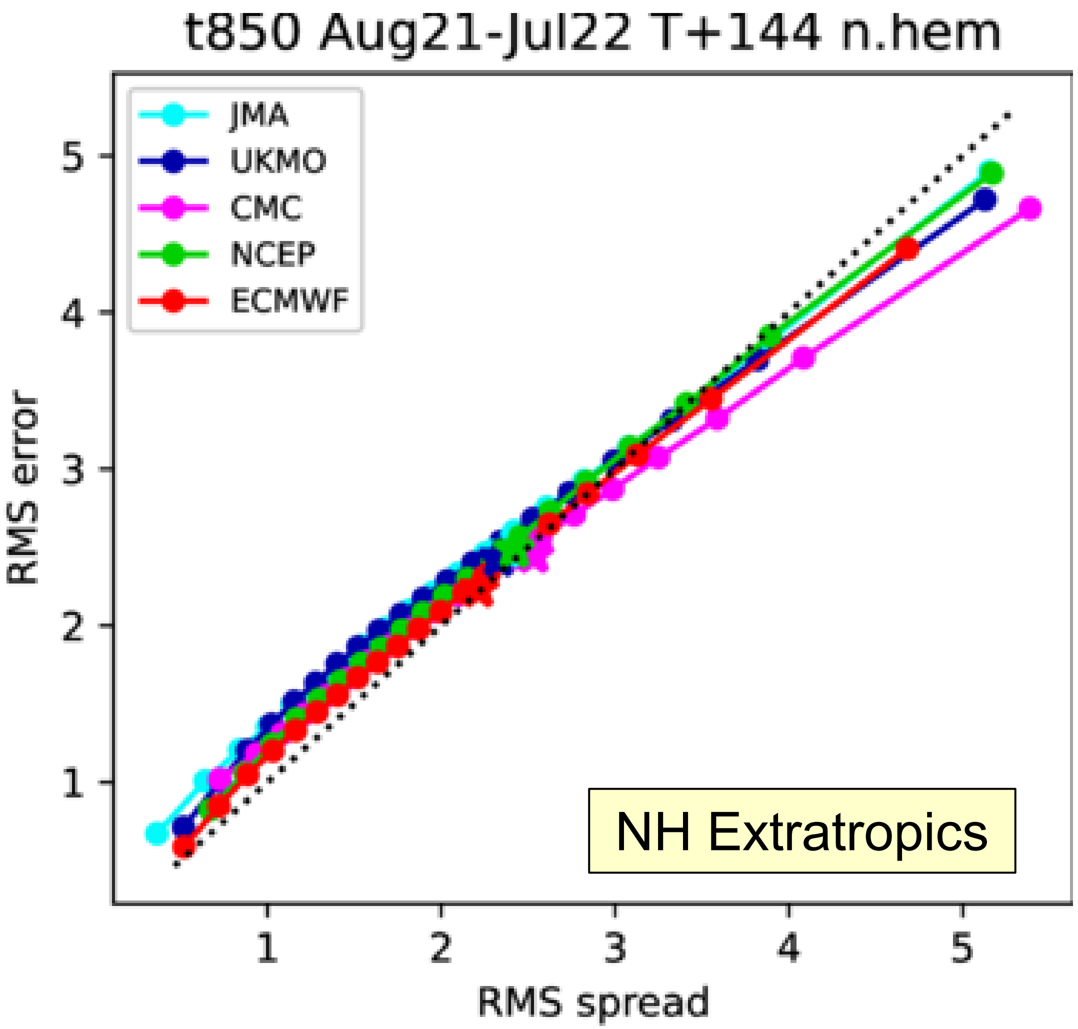
ENS: Z500 spread and error

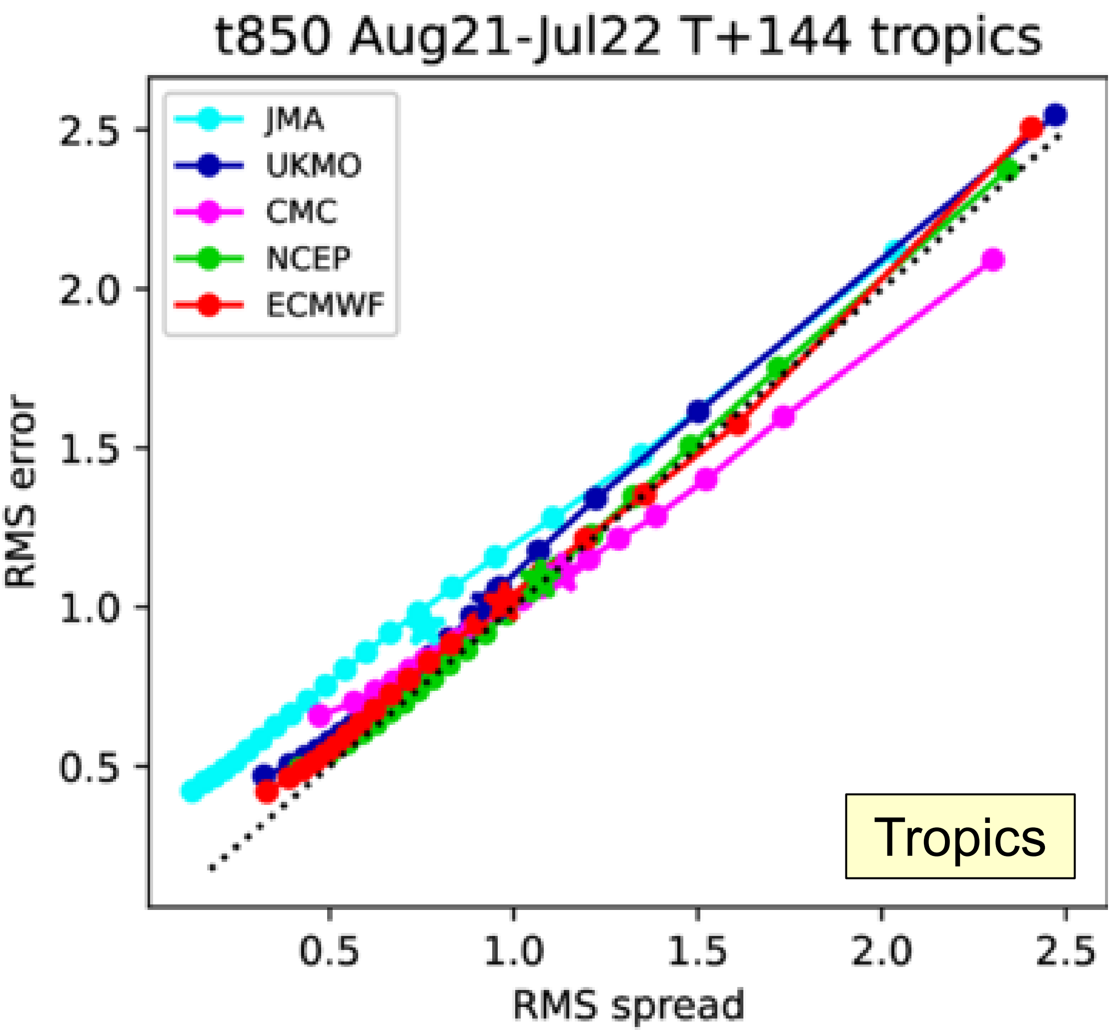


ENS: Z500 spread and error



T850: Spread reliability


t850 Aug21-Jul22 T+144 n.hem
JMA
UKMO
CMC
NCEP
ECMWF
RMS error
RMS spread
NH Extratropics


t850 Aug21-Jul22 T+144 tropics
JMA
UKMO
CMC
NCEP
ECMWF
RMS error
RMS spread
Tropics