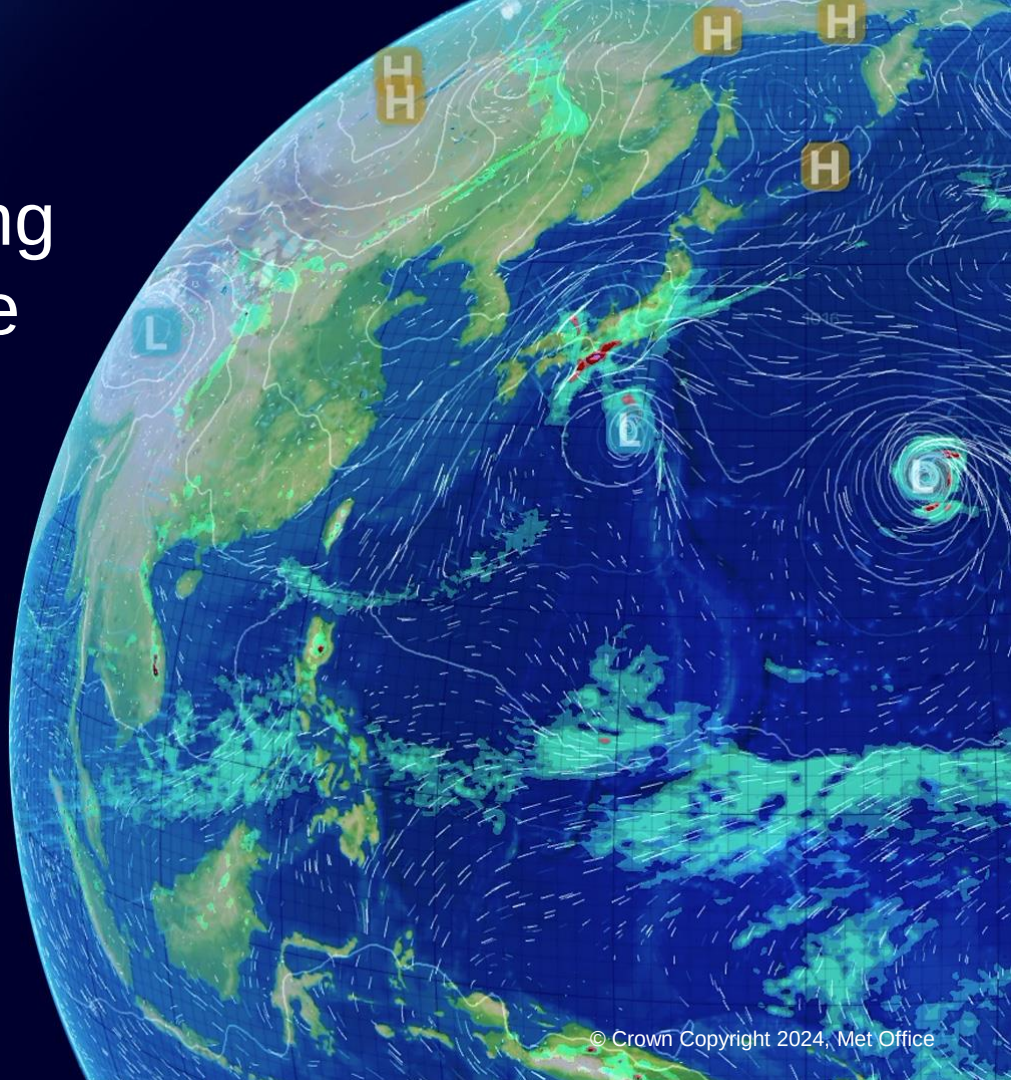
A framework for comparing post-processed percentile forecasts to a raw NWP ensemble

Roger Harbord, Laurence Beard, Clare Bysouth, Robert Coulson, Gavin Evans, Phil Gill, Simon Jackson, Chris Steele

Met Office, UK

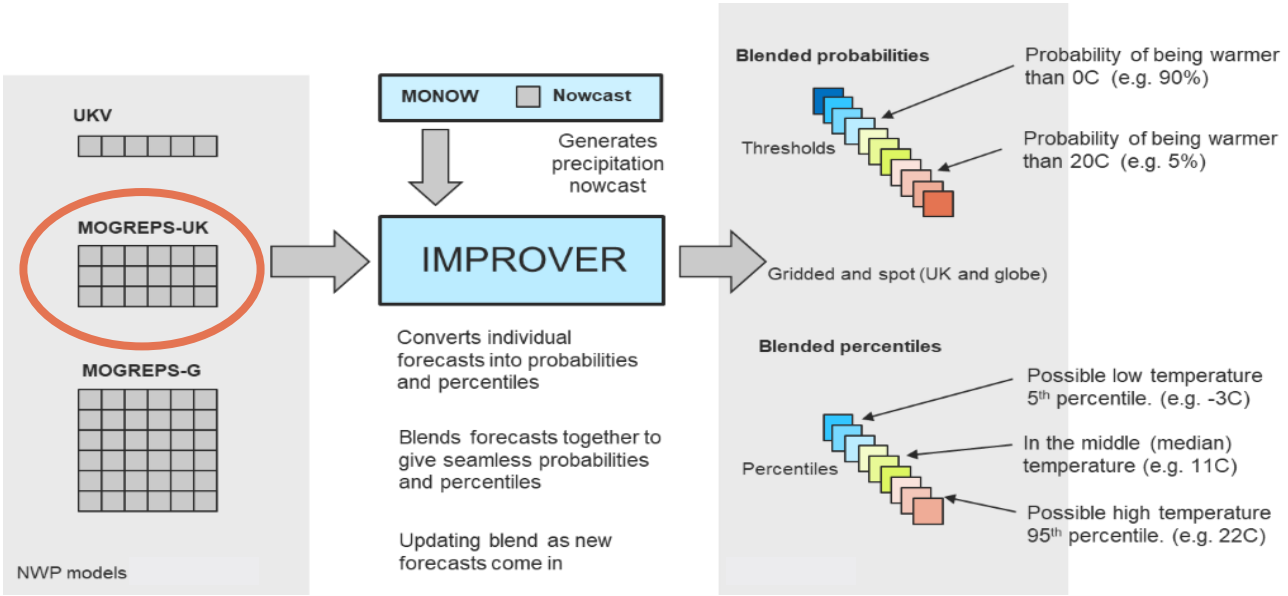
9th International Verification Methods Workshop, 21 May 2024



IMPROVER project: **I**ntegrated **M**odel post-**P**rocessing and **Ver**ification

Project completed earlier this year and the system is now operational.
Output available as the **Met Office Blended Probabilistic Forecast**

MOGREPS-UK:
18-member
regional ensemble



Roberts *et al.* 2023: "IMPROVER: The new probabilistic postprocessing system at the Met Office." *BAMS* 104: E680–E697

Principle objective:

Produce a scorecard summarising the increase in forecast accuracy over the UK provided by post-processing compared to the raw output from MOGREPS-UK for use as a key performance indicator (KPI)

Near-surface weather parameters

- 2-metre temperature
- 1-metre wind speed
- Precip accumulation (1-hour and 3-hour)
- Total cloud cover
- Visibility

Verified against 117 quality-controlled surface stations in the UK

Forecast ranges

Nowcast, short-range and medium-range up to T+120:

T+1, 2, 3, 6, 12, 24, 36, 48, 72, 96, 120

Metrics

- Ranked probability score (RPS)
- Continuous ranked probability score (CRPS)

Verification data storage and management

- Forecasts and matched observations stored in an Oracle database
- Over 2 terabytes of data in total so far
  - Over a year of hourly datetimes
  - Hourly lead times to T+120
  - 117 stations
  - 6 weather parameters
- Accessed from our Python-based verification front-end system using an SQL library (`sqlalchemy`)

MOGREPS-UK ensemble

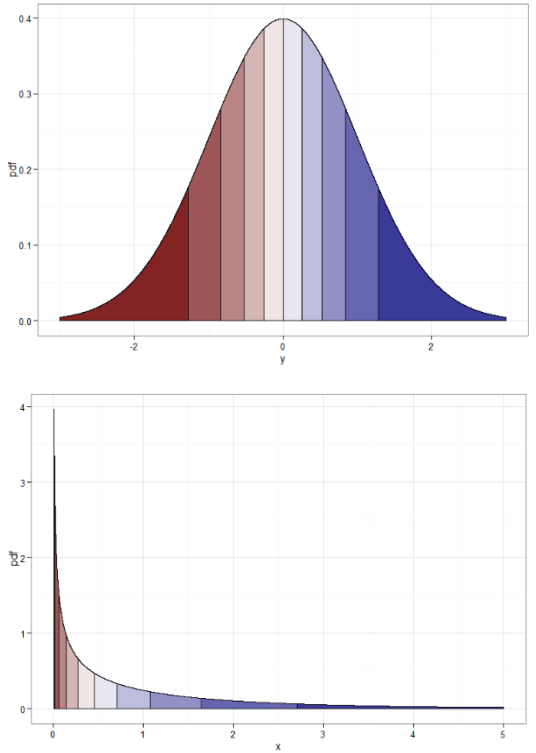
Date and time	Station ID	Lead time	Parameter	Obs	*Member* 0	*Member* 1	…

IMPROVER percentiles

Date and time	Station ID	Lead time	Parameter	Obs	5th *percentile*	10th *percentile*	…

Percentile forecasts

- IMPROVER outputs both percentiles and threshold exceedance probabilities
- Every 5th percentile:
  5%, 10%, 15%, … , 90%, 95%
  (*ventiles, vigintiles, or demi-deciles*)
- Many more than 20 thresholds for some variables
  - Temperature: intervals of 1°C between -20°C and 50°C
- Percentiles are more efficient and simpler to store
  - Typically more closely-spaced so contain more information
  - Simpler to store as there are the same number for all variables



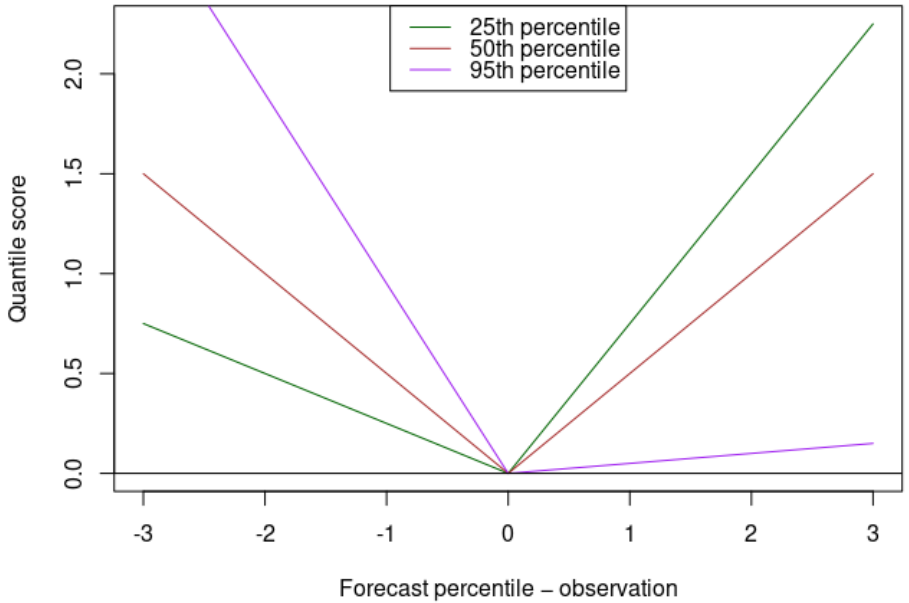
Deciles of normal and χ^2_1 distributions

The quantile score

The **quantile score** is a proper score for a quantile or percentile forecast. For an observation y and a forecast x_α for the α-quantile,

$$\mathrm{QS} = \begin{cases} (x_\alpha - y)(1-\alpha), & y \le x_\alpha \\ (y - x_\alpha)\alpha, & y > x_\alpha \end{cases}$$

- For the median (50th percentile), the QS is equal to half the mean absolute error
- The QS is zero for the 0% and 100% quantiles provided the observation lies in between them.



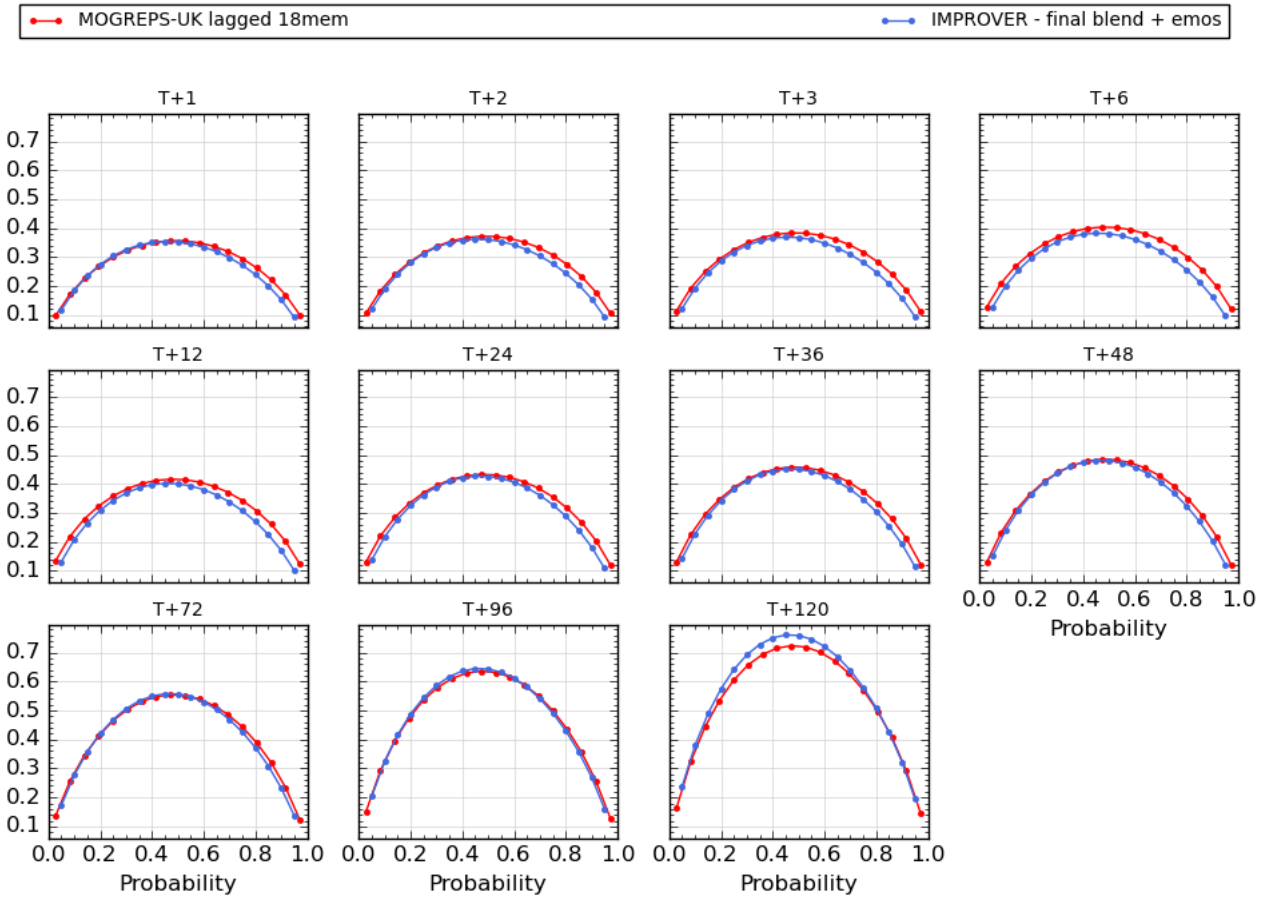
CRPS from quantile scores

- **For a continuous distribution**, the CRPS is equal to twice the integral of the quantile score over the quantile level α.
- **For an ensemble**, Bröcker (2012) shows that the CRPS can be written as the mean of the quantile scores applied to the individual members, using level $(k - 0.5)/M$ for the k^{th} of M ensemble members.
  - So with M = 19, the appropriate quantile levels are 0.026, 0.079, 0.132, … 0.974

Bröcker, 2012: "Evaluating raw ensembles with the continuous ranked probability score." *QJRMS* 138: 1611-1617.

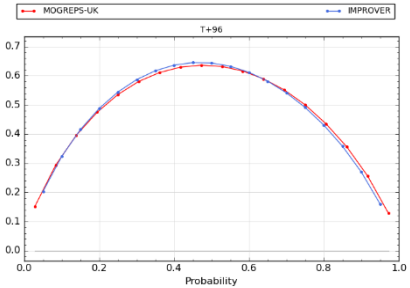
Quantile scores for temperature

- Up to T+72, IMPROVER has a better score than MOGREPS-UK at higher quantiles
- At T+120, IMPROVER appears worse around the middle of the distribution

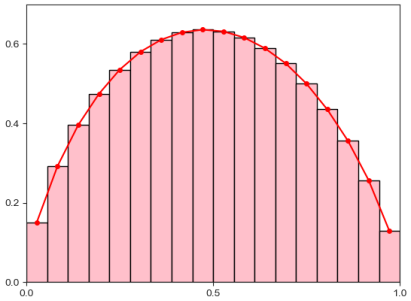


How to calculate CRPS from a set of discrete quantiles?

5%, 10%, 15%, 20%, … 95%

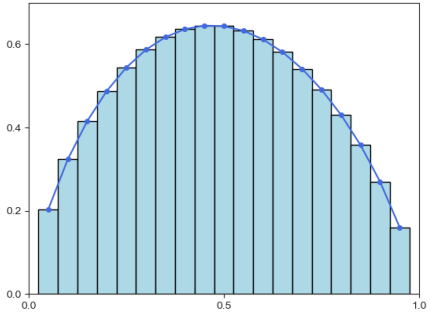


MOGREPS-UK

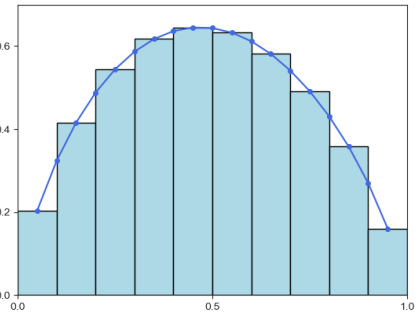


Bröcker's instructions are equivalent to the midpoint rule

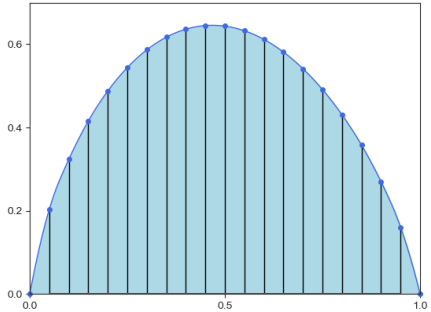
IMPROVER – which is the most appropriate?



Average the quantile scores for the 19 percentiles



Midpoint rule using only 5%, 15%, 25%, … 95%



Simpson's rule

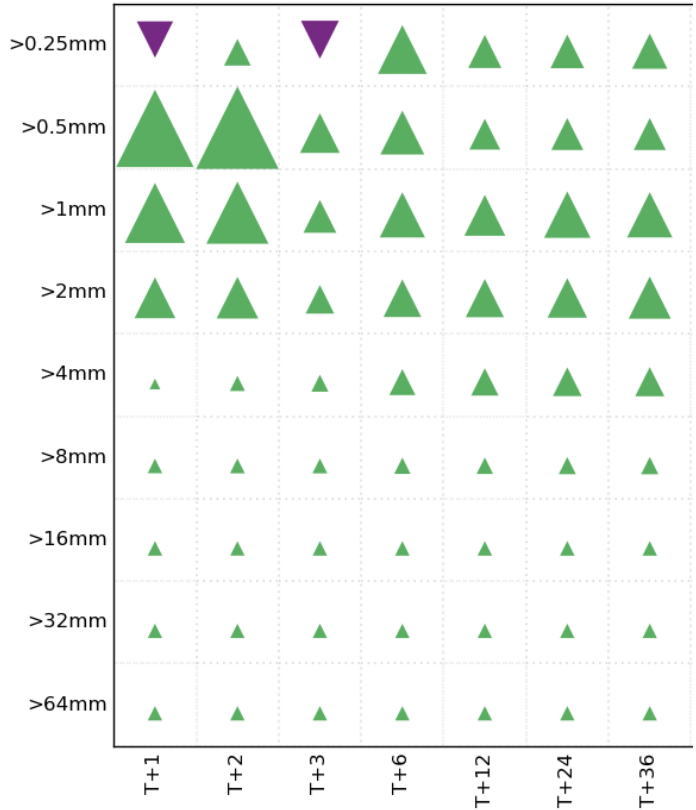
RPS and Brier scores

The RPS is the sum of Brier scores across thresholds.

It seemed reasonable to assign probabilities:

- 0.075, 0.125, …0.925 to thresholds lying between IMPROVER's percentiles
  0.025 (0.975) to thresholds lying below the 5th (above the 95th) percentile
- $(k + 0.5)/19$ for $k = 0…18$ to MOGREPS-UK ensemble member counts, giving 0.026, 0.079, 0.132, … 0.974

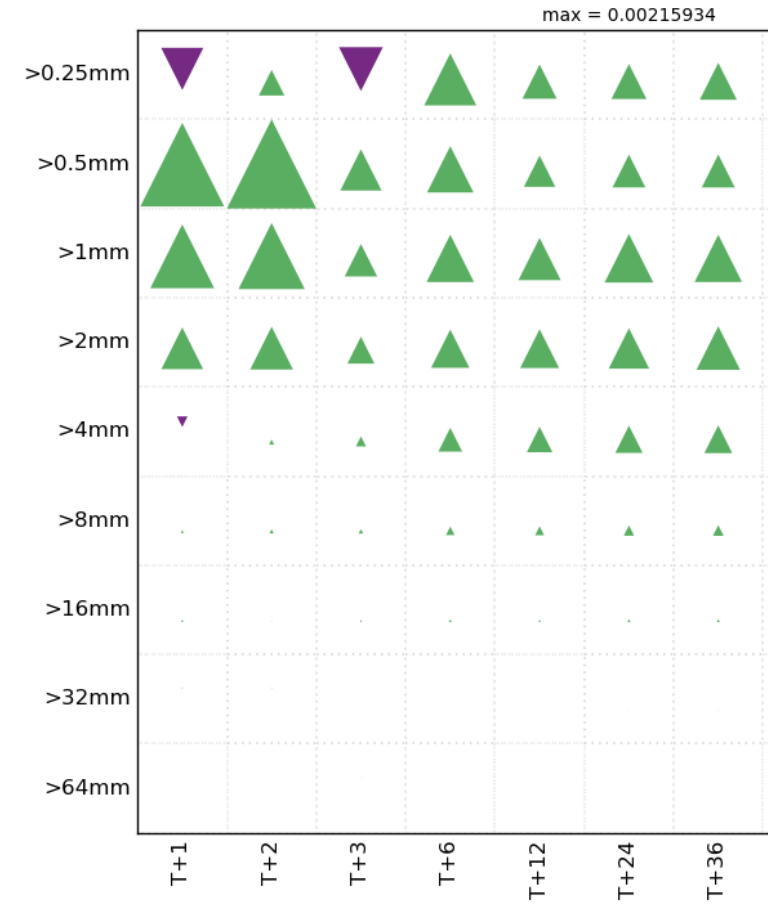
⇒ IMPROVER scores better for extreme categories, including those that were never observed and were forecast with the smallest possible probability for both systems!



Differences in Brier scores between the two systems

A fix

- One alternative is to change the assignment of probabilities to MOGREPS-UK member counts.
- In general for an *M*-member ensemble, you can assign probabilities $(k + a)/(M + 2a)$ for a chosen value of a between 0 and 1.
  (Wilks, 2019: *Statistical Methods in the Atmospheric Sciences* sections 8.3.1 and 3.3.7)
- Choosing $a = 9/19$ instead of 1/2 results in the smallest probability that can be issued being the same for both systems, namely 0.025
- This gives zero difference between the Brier scores for the two systems in the extreme categories.

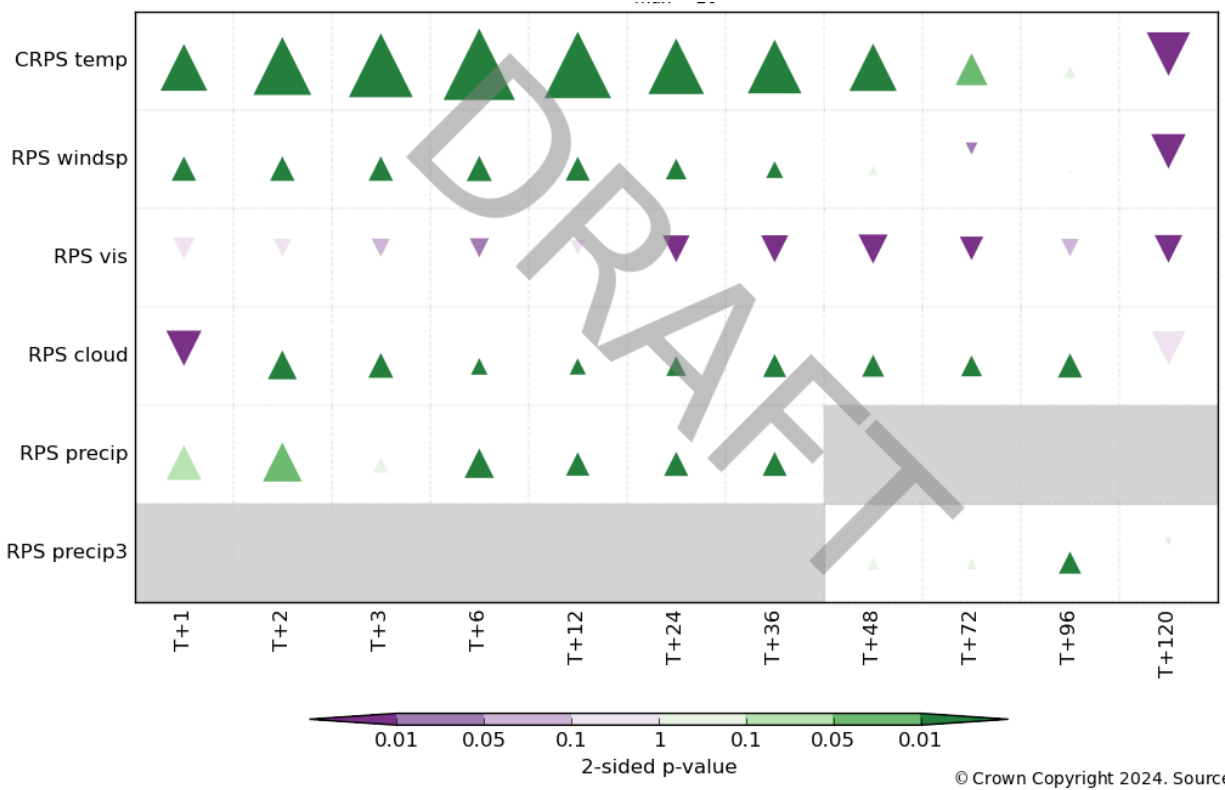


plot source

Summary scorecard

- Size of triangle indicates magnitude of the percentage difference
- Orientation and colour indicates sign of difference
- Shade of colour indicates p-value (adjusted for serial correlation using AR(1) variance inflation applied to daily means of score differences)

Scores averaged over April 2023 to March 2024



Summary

- Storing post-processed spot forecasts as percentiles allows flexible fully probabilistic verification and comparison to an ensemble
- Quantile scores give insight into benefits of post-processing at extremes
- Calculation of CRPS, Brier scores and RPS needs some care to ensure a fair comparison
- The IMPROVER project has successfully delivered a new Met Office blended probabilistic forecast

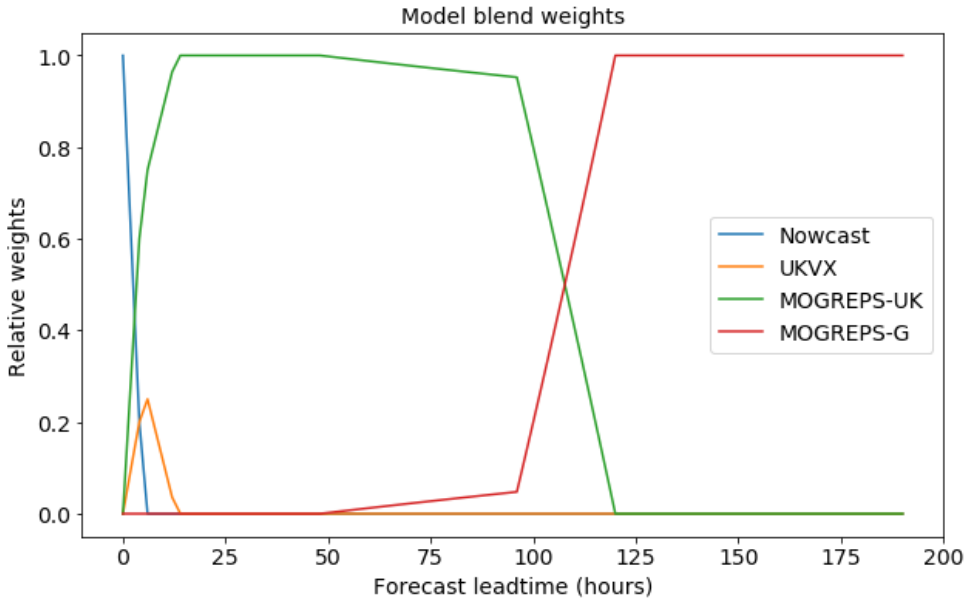
  … and verification will be essential to its continued improvement

roger.harbord@metoffice.gov.uk

Extra slides

Model blending

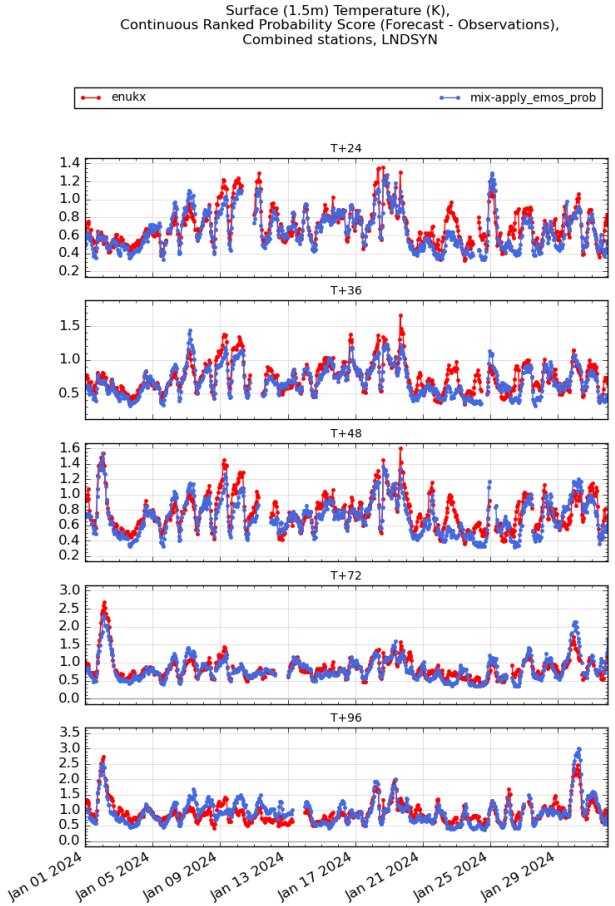
- MOGREPS-UK dominates the weighting at T+6 to T+96, and currently runs out to T+120.
- Its weight tapers to zero at T+120 to give a smooth transition to the global ensemble MOGREPS-G



Beyond the scorecard

e.g. monitoring CRPS for temperature over Jan 2024

- Sometimes both systems do poorly (large CRPS) e.g. around 19th
- Sometimes MOGREPS-UK has a large peak where IMPROVER doesn't, e.g. 23rd



plot source

IMPROVER code is open-source: improver.readthedocs.io