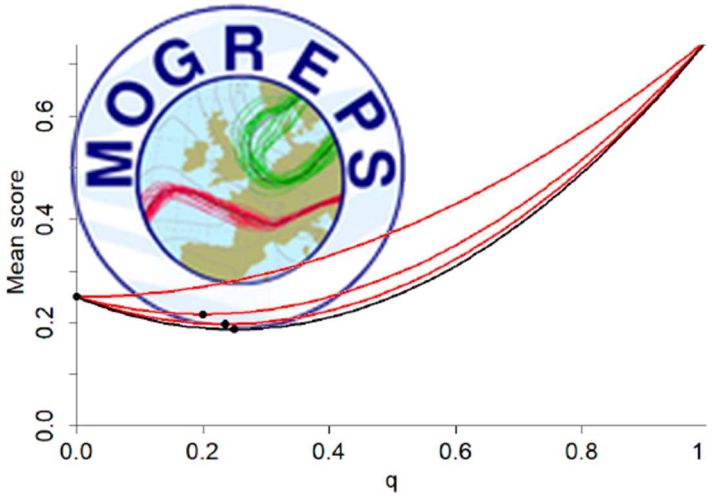
Application of fair scores to lagged and unlagged ensembles from hourly-cycling MOGREPS-UK

Roger Harbord roger.harbord@metoffice.gov.uk

45th EWGLAM and 30th SRNWP meeting
27 Sep 2023



Acknowledgements

- Chris Ferro, University of Exeter

At the Met Office:

- Rob Darvell, Teresa Hughes, Anette Van der Wal, Clare Bysouth, Phil Gill, Ric Crocker, Jo Robbins
- Nigel Roberts, Marion Mittermaier

This work was conducted through the Weather and Climate Science for Service Partnership (WCSSP) India, a collaborative initiative between the Met Office, supported by the UK Government's Newton Fund, and the Indian Ministry of Earth Sciences (MoES).

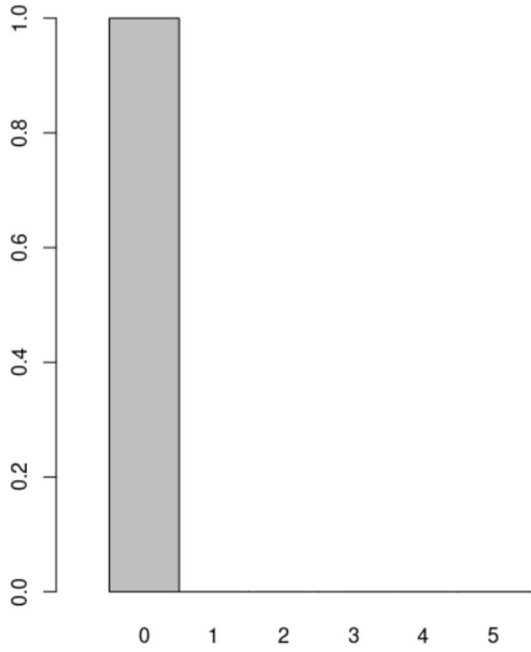
Contents

- Realistic ensembles and fair scores
- Definition of the fair Brier score
- Application to unlagged and lagged hourly cycling MOGREPS-UK
- Implications and discussion

Fair scores for ensemble forecasts

Chris Ferro, *QJRMS* 2014 DOI:10.1002/qj.2270

- *Proper* scores are appropriate for assessing forecasts issued as probability distributions (Brier score, RPS, CRPS…)
- But ensembles forecasts are *not* probability distributions – they should behave like *a random sample* drawn from the same probability distribution as the verifying observation
- If no members of an ensemble forecast contain an event, we cannot be 100% certain that the event will not occur
- If an event occurs on 2.5% of occasions, an 18-member ensemble that never predicts the event will on average have a *better* Brier score than an ensemble sampled from a distribution with Prob(event) = 0.025



Realistic ensemble forecasts and fair scores

- Definition: An ensemble forecast is *realistic* if the ensemble members and outcome behave as if they are drawn from the same distribution.
- To assess ensemble forecasts for realism we should:
  - assess calibration with rank histograms
  - measure overall performance with *fair scoring rules*
- The expected value of *fair* scores is optimised for *realistic* forecasts

Conventional (original, unadjusted) Brier score

- k out of M ensemble members predict the binary outcome y will occur
- Let $p_k = k/M$
  (The proportion of members predicting the outcome to occur)
- The conventional Brier score is

$$\mathrm{BS} = (p_k - y)^2$$

For a 3-member ensemble:

		Outcome	
k	p_k	$y = 0$	$y = 1$
0	0	0	1
1	1/3	1/9	4/9
2	2/3	4/9	1/9
3	1	1	0

Fair Brier score

$$\text{Fair BS} = (p_k - y)^2 - \frac{p_k(1-p_k)}{M-1}$$

Ensembles for which all but one member makes a correct forecast score the same as perfect forecasts
(Ferro, *QJRMS* 2014)

For a 3-member ensemble:

		Outcome	
k	p_k	*y = 0*	*y = 1*
0	0	0	1
1	1/3	0	1/3
2	2/3	1/3	0
3	1	1	0

Hourly cycling MOGREPS-UK

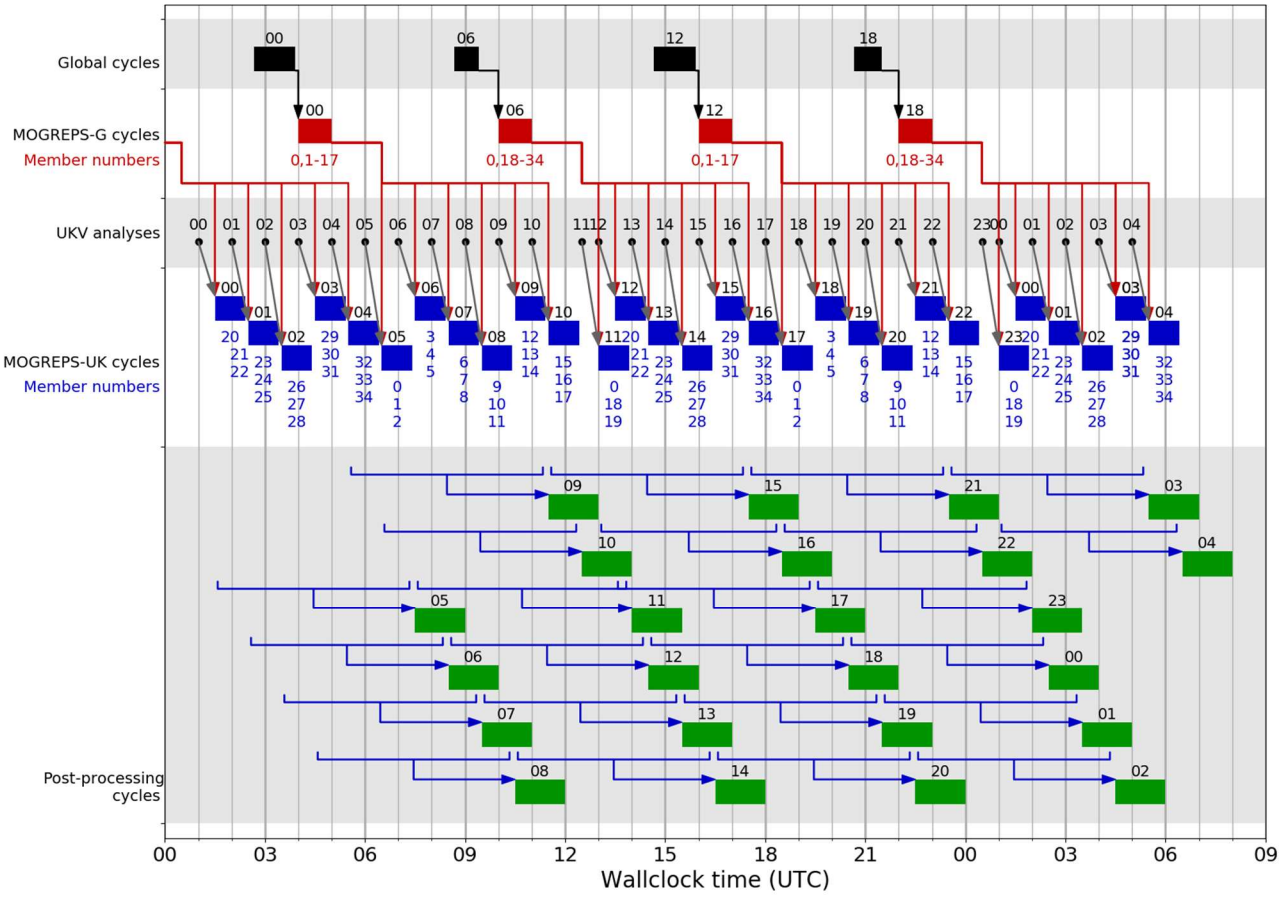
Porson, Carr, Hagelin *et al.* (2020)

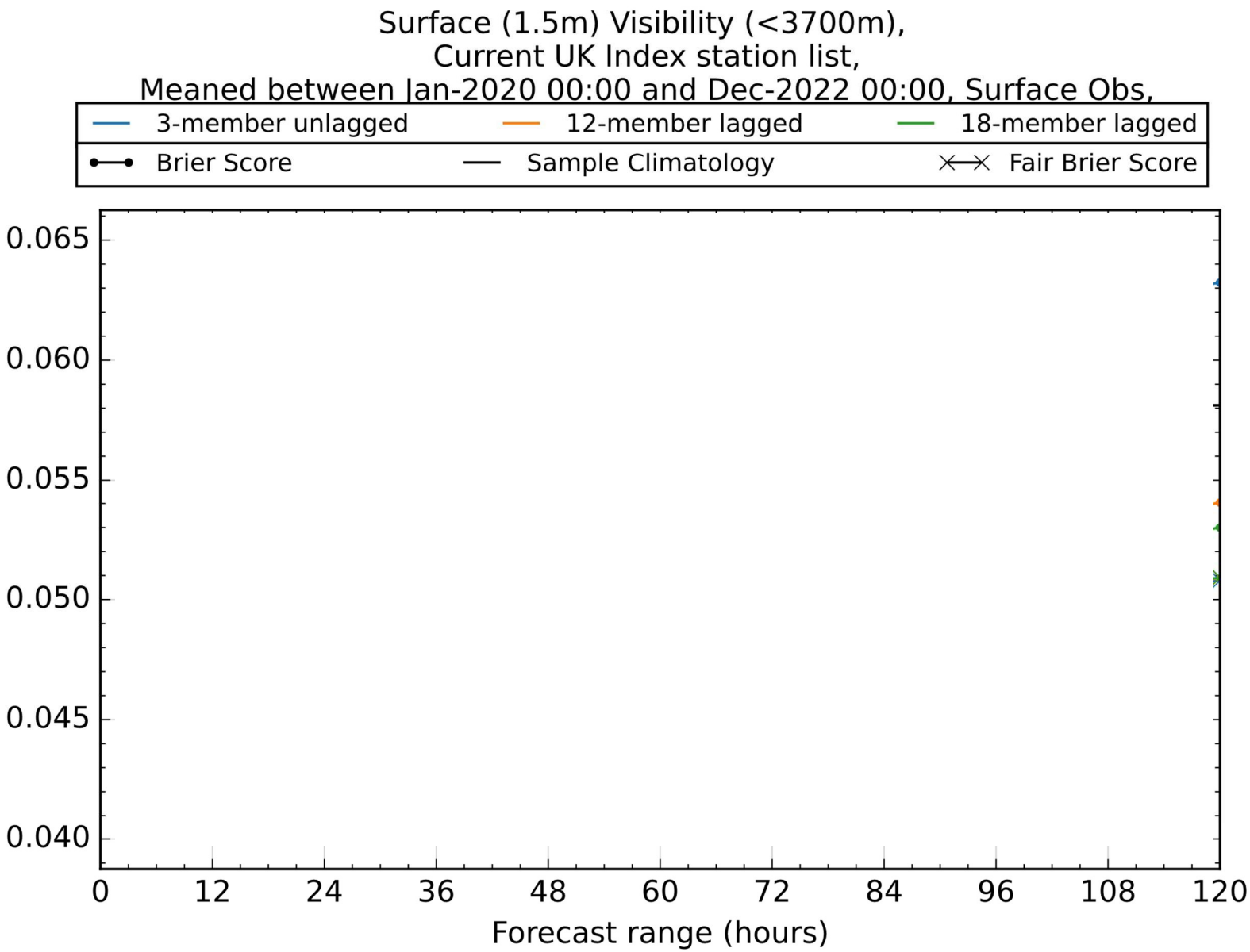
- *05, 11, 17, 23 UTC cycles*:
  1 control run + 2 perturbed members
- *All other cycles*:
  3 perturbed members

(Since Dec 2019)

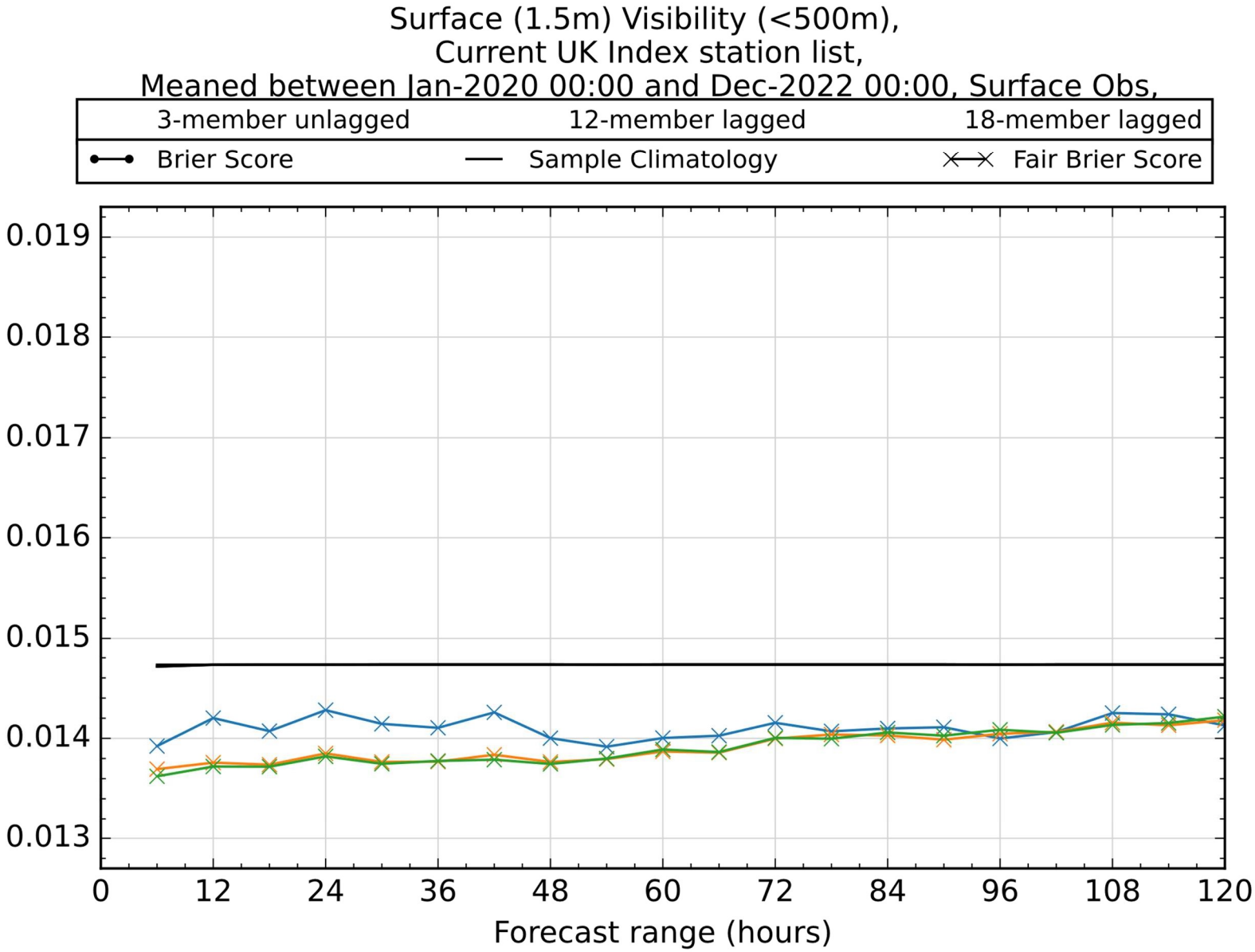
An 18-member ensemble is created by time-lagging over the 6 most recent cycles.

(or a 12-member ensemble by time-lagging over 4 recent cycles)





Brier scores for visibility below 3700 metres



Brier scores for visibility below **500** metres

… when all 3-member ensembles *include* a control member

Other surface parameters

Wind speed, one-hour precipitation accumulation, cloud cover, cloud base height:

- Many parameters and thresholds show smaller or no difference between the fair Brier scores for the unlagged and lagged ensembles even at the shortest forecast ranges
- The fair scores for the *lagged* ensembles are *never worse* than for the unlagged

HiRA

- HiRA was designed to overcome the double-penalty effect:
  - ✓ When assessing high-resolution deterministic forecasts
  - ✓ When comparing deterministic forecasts to ensembles

But is it the best method for assessing ensembles by themselves?

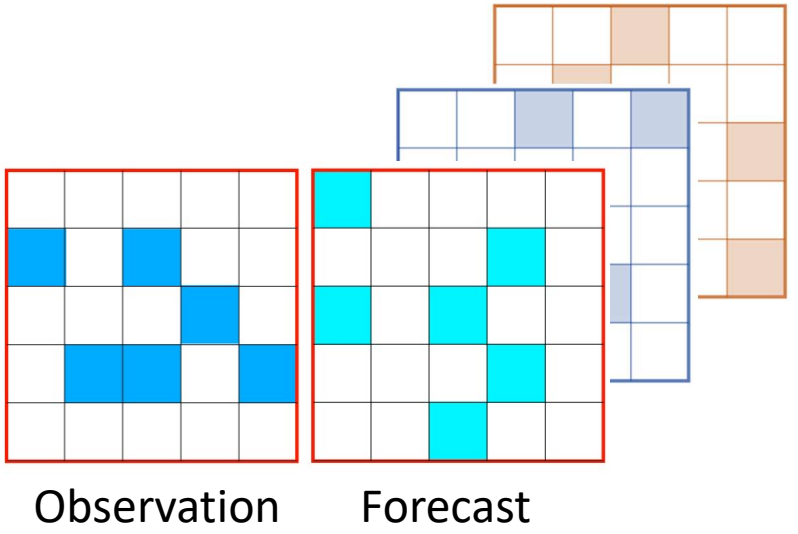


Illustration of the double-penalty effect

HiRA and fair scores

- HiRA generates a pseudo-ensemble from forecasts in the neighbourhood of each gridpoint.
- The double-penalty effect for ensembles *is essentially the same thing* as the issue with unadjusted scores being worse for realistic ensembles
- There is no fair score for an ensemble of size one, just as there is no way of overcoming the double-penalty effect for a deterministic forecast using gridpoint scores.
- In HiRA, the double-penalty effect decreases with neighbourhood size, just as the difference between unadjusted and fair scores does.
- I contend that *fair scores provide a simpler and better way* of 'extrapolating the ensemble size to infinity'

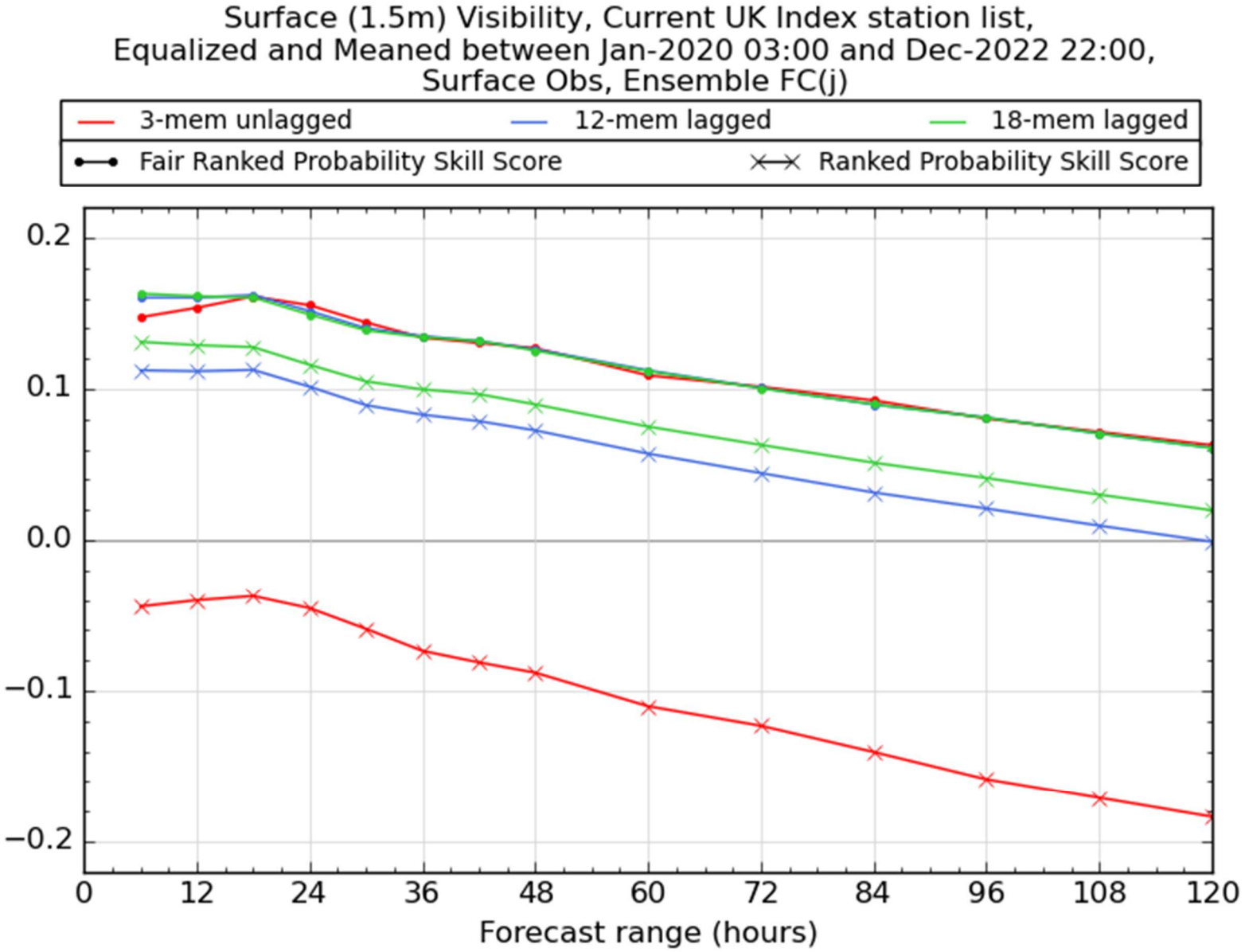
Summary

- Fair scores are used to assess whether modest-sized ensembles behave *realistically*, particularly with rare events
- The theory of fair scores applies surprisingly well to MOGREPS-UK
- The results help to show that time-lagging hourly MOGREPS-UK works well

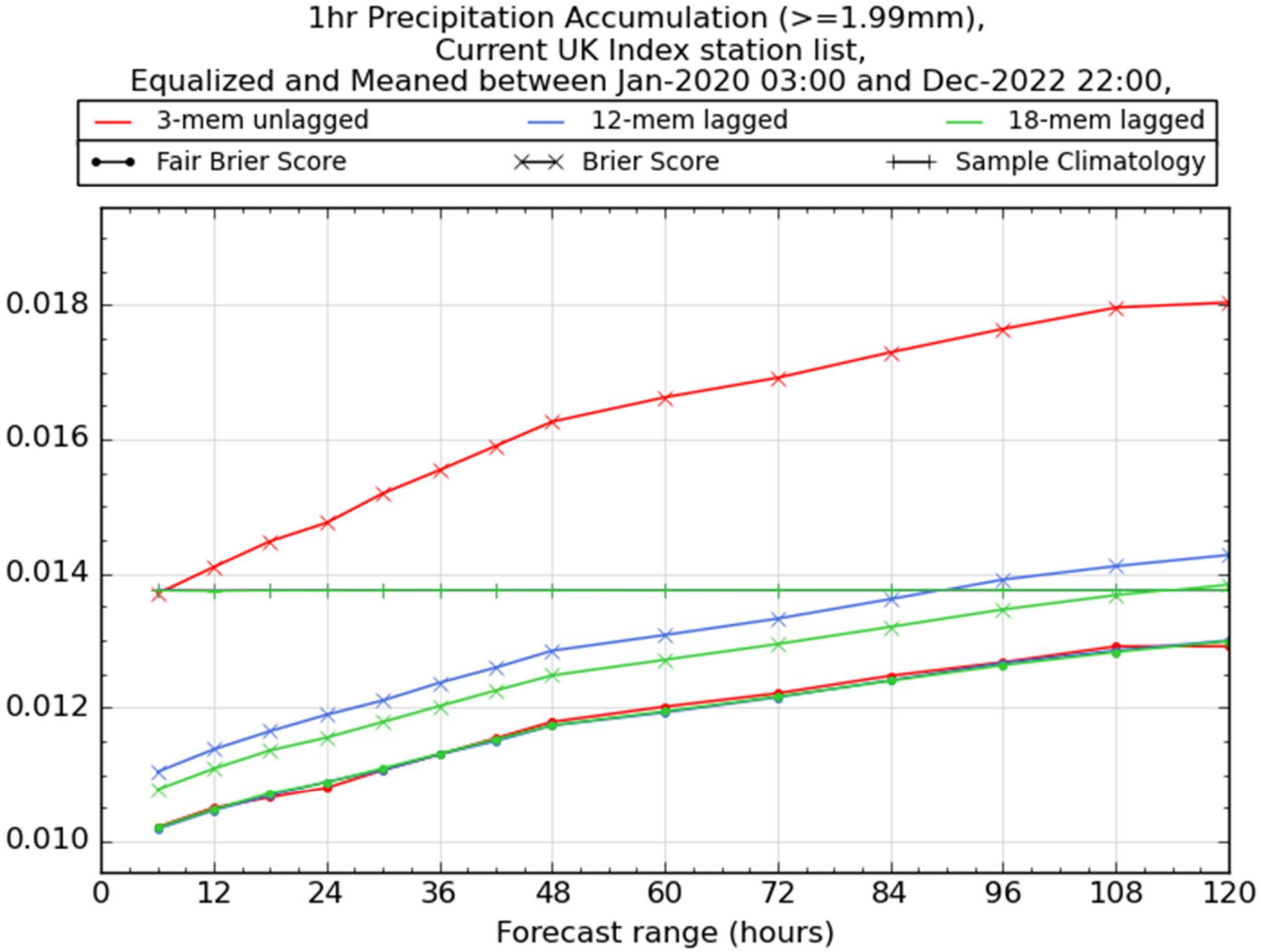
- Fair scores can help clarify and simplify verification of ensemble systems and trials

roger.harbord@metoffice.gov.uk

Supplementary slides



Fair and unadjusted **RPSS** for visibility



Fair and unadjusted Brier scores for **1-hr precip accumulation > 1.99 mm**

[Link to plots of other thresholds and parameters]