A diagram to assess consistency of ensemble forecasts of binary events
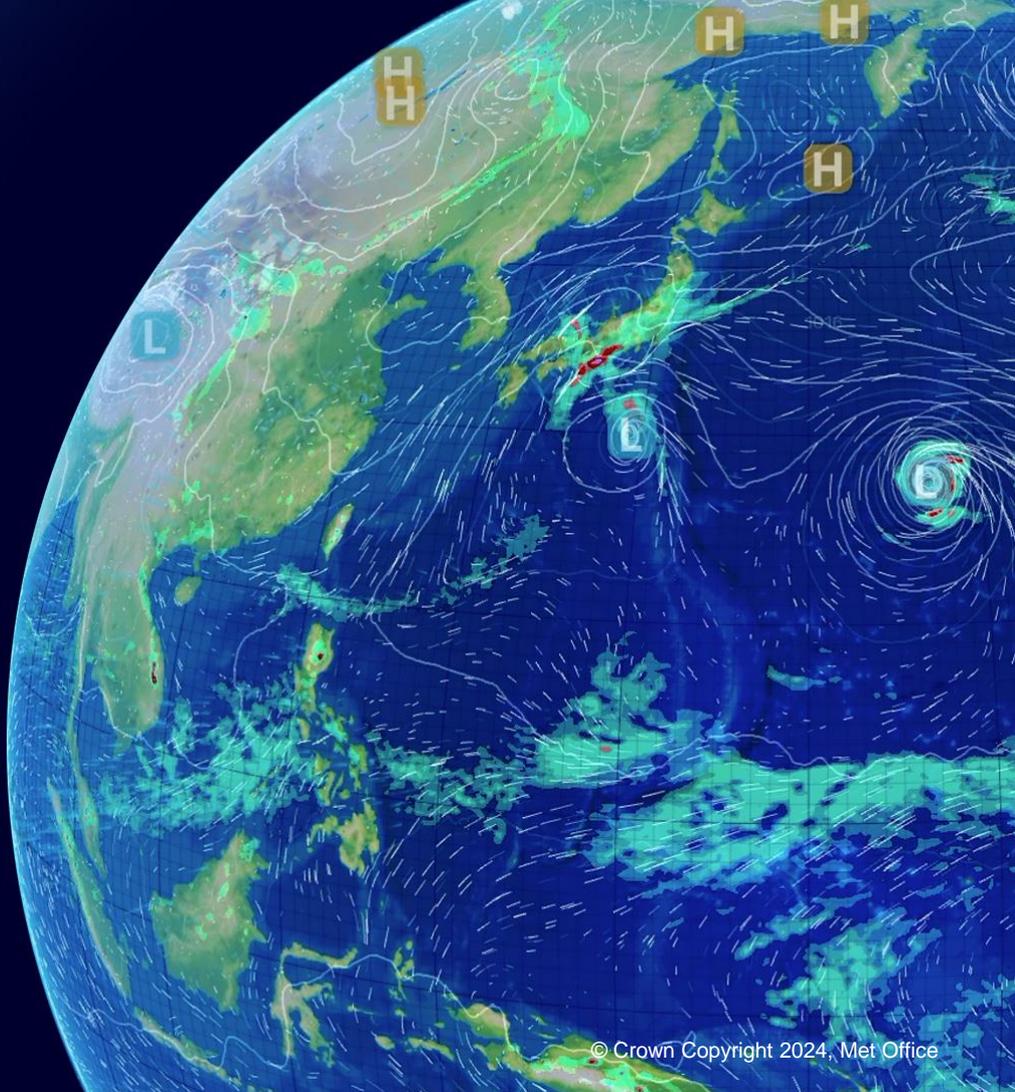
Roger Harbord

Met Office, UK

roger.harbord@metoffice.gov.uk

46th EWGLAM & 31th SRNWP Meeting

30 September 2024

Reliability (calibration) diagrams for *probabilistic* forecasts

For a binary event:

- Group the probability forecasts into bins

- Plot the proportion of times the event occurred when the forecast fell into each bin against a typical probability for that bin.
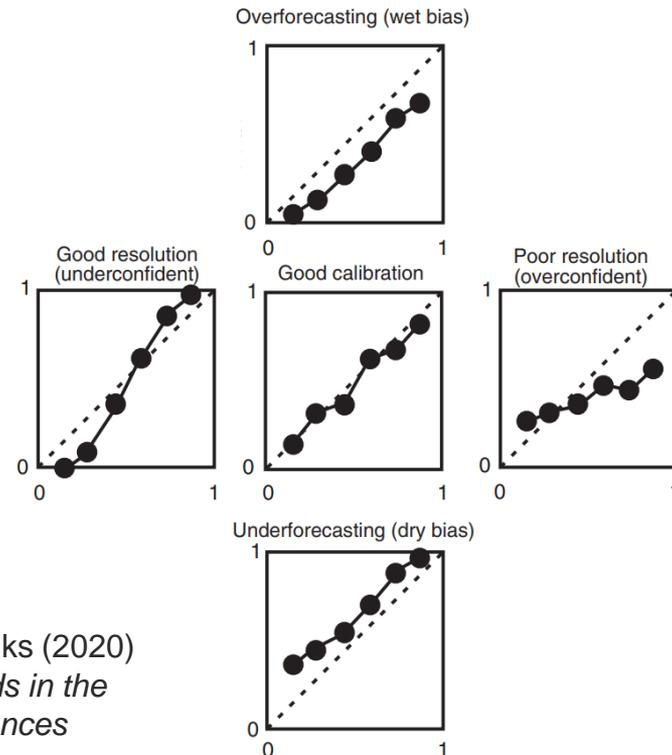

Figure 9.10 of Wilks (2020)
Statistical Methods in the Atmospheric Sciences

Ensemble consistency (reliability)

Definition:

An ensemble forecast is *consistent*, or *reliable*, if the ensemble members and the verifying observation behave as if they have been sampled from the same distribution.

Fair scoring rules measure ensemble performance in a way that favours ensembles that are reliable in this sense

See Ferro, 2014 "Fair scores for ensemble forecasts" *QJRMS.* DOI:10.1002/qj.2270

Reliability diagrams for *ensemble* forecasts

When probability forecasts are derived from the fraction of ensemble members that predict the event, the forecasts appear overconfident when plotted on a reliability diagram *even if the ensemble is perfectly reliable*.
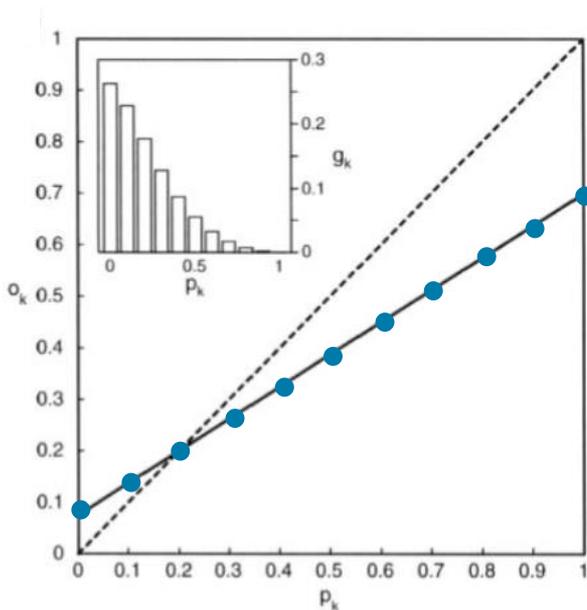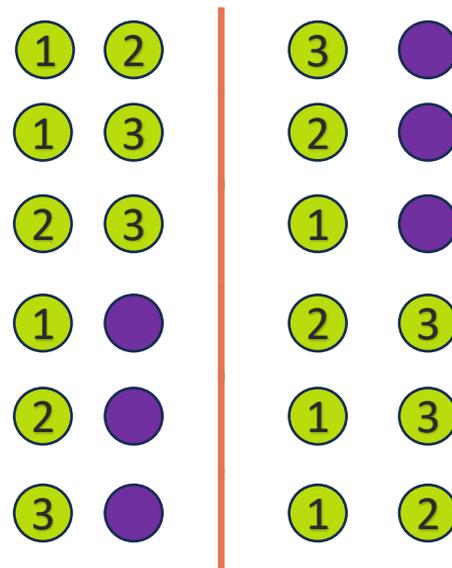
DS Richardson, 2001:
"Measures of skill and value of ensemble prediction systems, their interrelationship and the effect of ensemble size."
QJRMS 127: 2473-2489

An alternative

- Consider a three-member ensemble forecast *plus its verifying observation* as a single set with four elements.

- If the ensemble is reliable and we have a large number of such sets, of all the sets in which 2 of the 4 elements exceed the threshold, we expect the *observation* to exceed the threshold in half the sets.

- Similarly for 1 and for 3 elements exceeding the threshold.

- Plotting the *actual* fraction of the sets in which the observation exceeds the threshold against these *expected* fractions gives an 'ensemble consistency diagram'

Similar 'trick' to Bröcker & Ben Bouallègue 2020 "Stratified rank histograms …" *QJRMS*. DOI:10.1002/qj.3778

Sketch of proof

Consider a single ensemble forecast consisting of binary variables $X_1, X_2, \ldots X_m$, and a verifying observation Y. Let $\sum X_i = K$. Condition on a specific value, j, of the sum of the ensemble members and the observation.

Clearly,

$$\mathsf{E}\left(Y + \sum X_i \,\middle|\, Y + \sum X_i = j\right) = j$$

$$\mathsf{E}(Y \mid Y + K = j) + \sum \mathsf{E}(X_i \mid Y + K = j) = j$$

If Y and all the X_i are exchangeable, $\mathsf{E}(Y \mid Y + K = j) = E(X_i \mid Y + K = j)$ for all i.

Therefore

$$(m + 1)\,\mathsf{E}(Y \mid Y + K = j) = j$$

$$\mathsf{E}(Y \mid Y + K = j) = \frac{j}{m + 1}$$

Conventional reliability table for 3-member ensemble

Number of ensemble members predicting event	0	1	2	3
Number of events observed	a_0	a_1	a_2	a_3
Number of non-events observed	b_0	b_1	b_2	b_3
Observed frequency	$a_0/(a_0 + b_0)$	$a_1/(a_1 + b_1)$	$a_2/(a_2 + b_2)$	$a_3/(a_3 + b_3)$

# 'Ensemble consistency table' for 3-member ensemble

Number of elements in {ens members + obs} with event	1	2	3
Number of events observed	a_0	a_1	a_2
Number of non-events observed	b_1	b_2	b_3
Observed frequency	$a_0/(a_0 + b_1)$	$a_1/(a_1 + b_2)$	$a_2/(a_2 + b_3)$

Example: Hourly-cycling MOGREPS-UK

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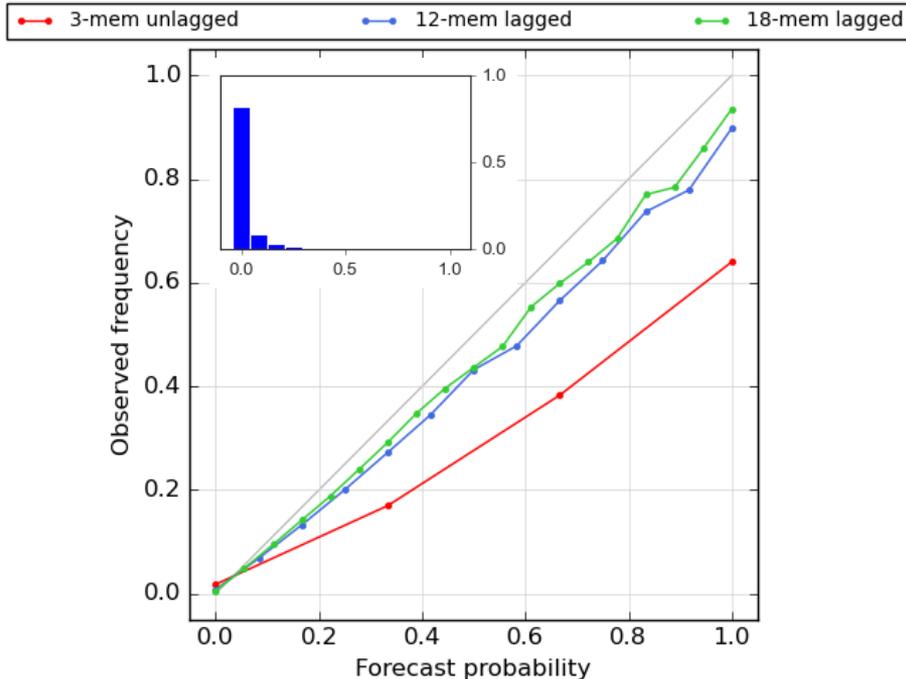
- *All other cycles*:
  3 perturbed members

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

Porson, Carr, Hagelin *et al.*, 2020. "Recent upgrades to the Met Office convective-scale ensemble: an hourly time-lagged 5-day ensemble." *QJRMS*. DOI:10.1002/qj.3844
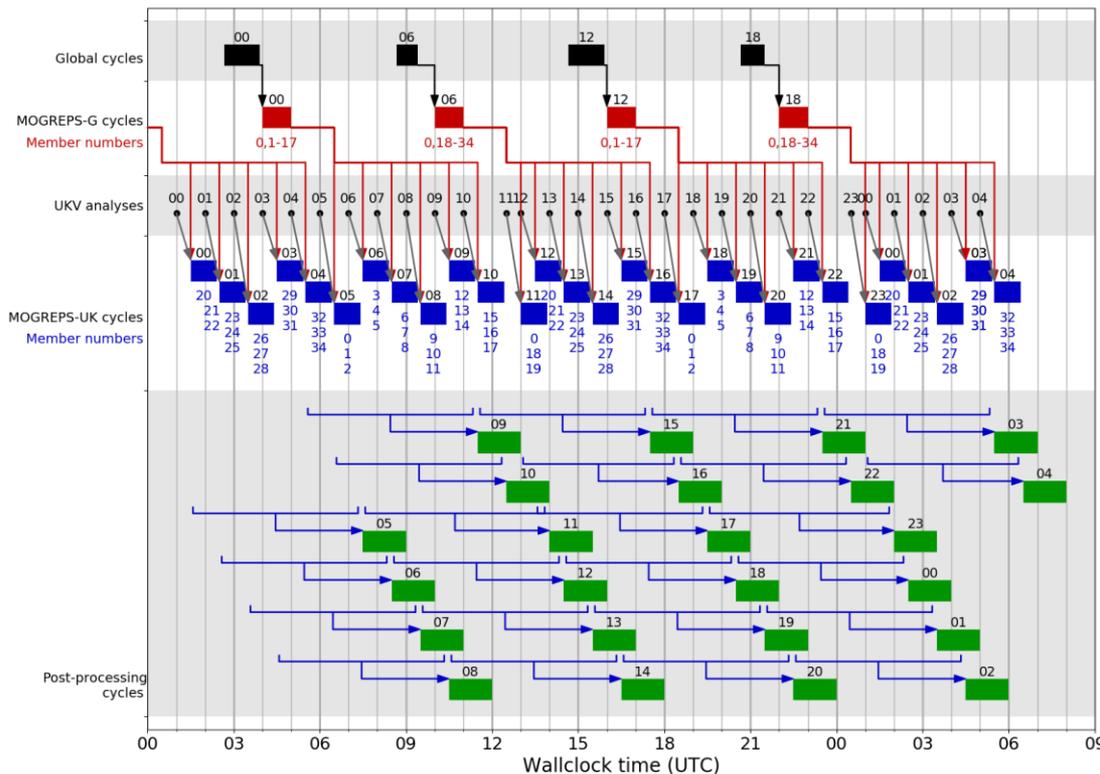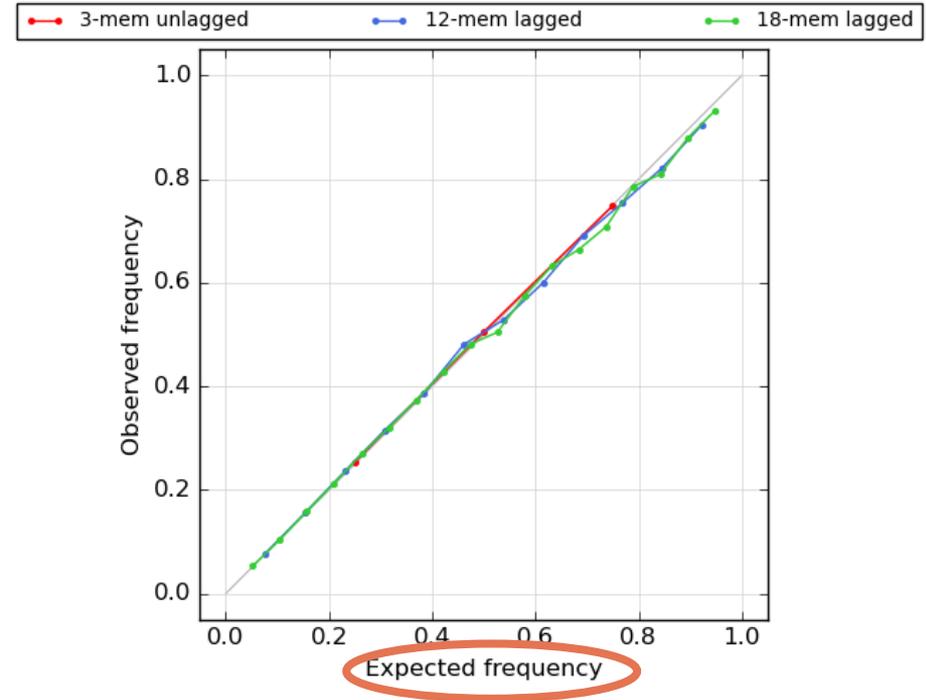
1-hr precip accumulation ⩾1 mm, T+24
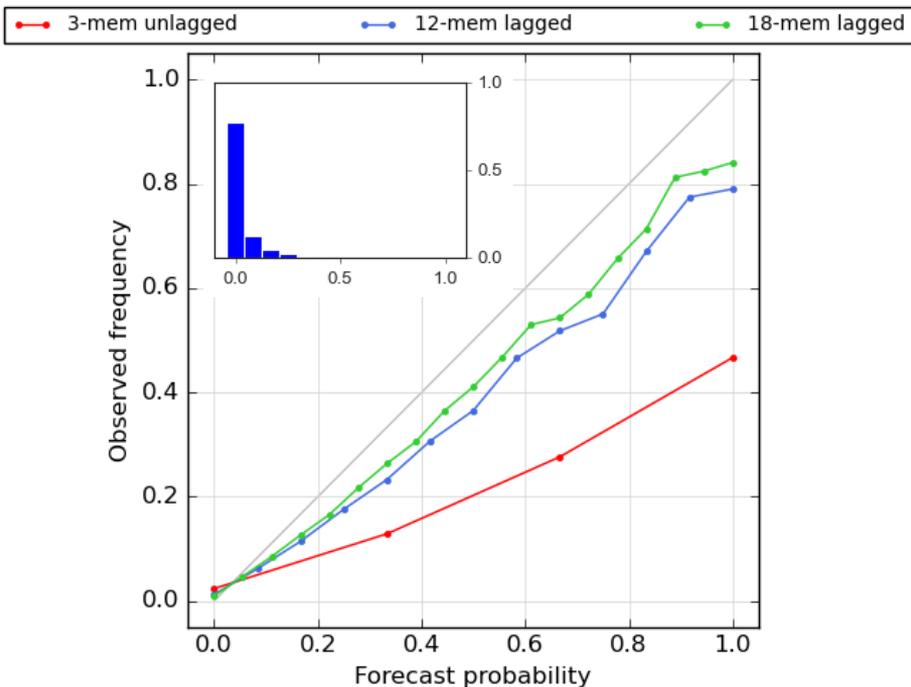

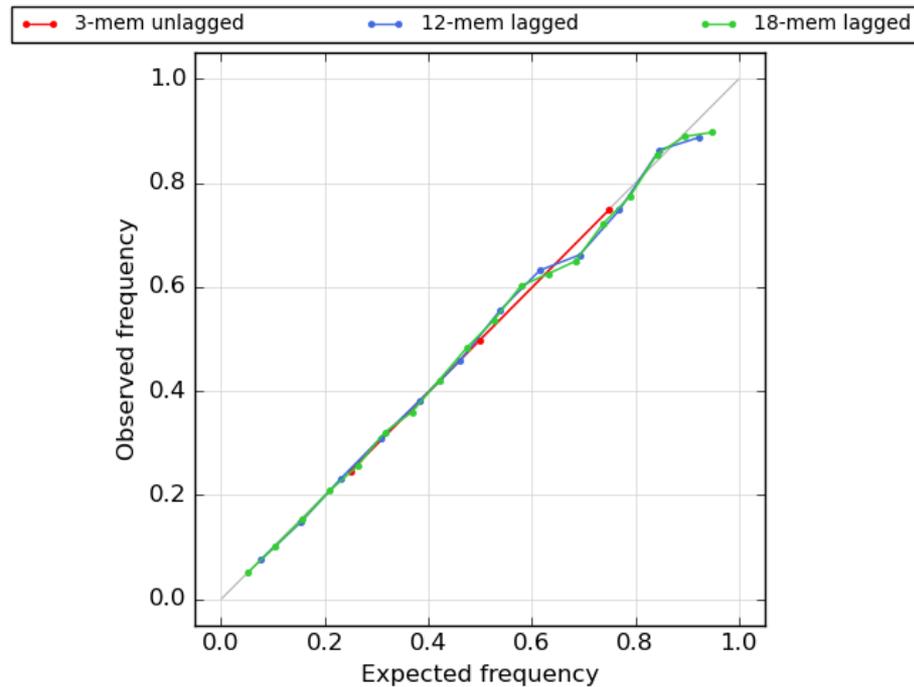Conventional reliability diagram

Ensemble consistency diagram

1-hr precip accumulation ⩾1 mm, **T+72**

Conventional reliability diagram

Ensemble consistency diagram
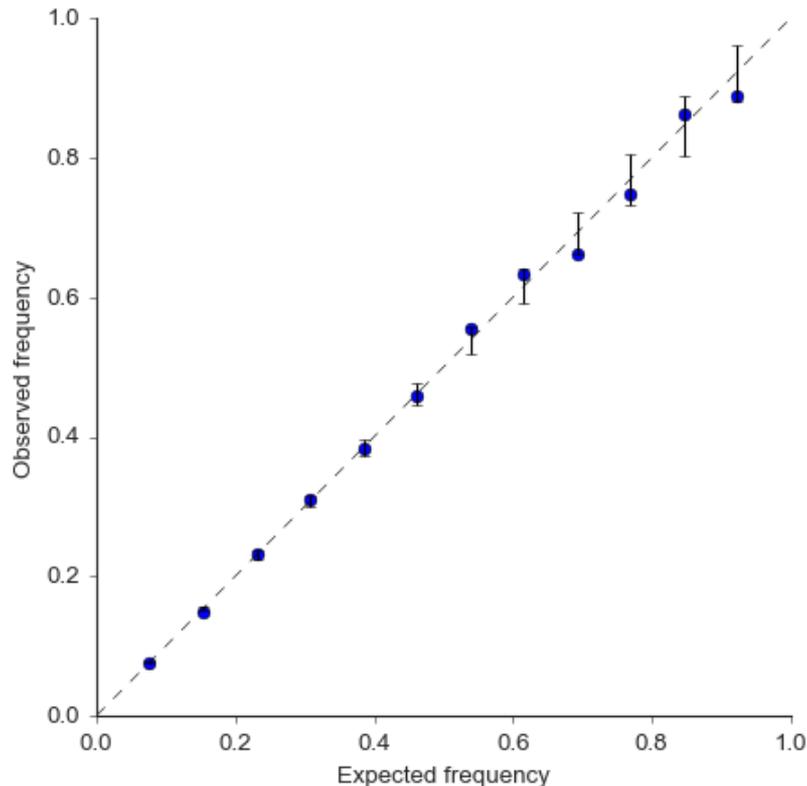
Consistency bars

Similar to Bröcker & Smith, 2007.
"Increasing the reliability of reliability diagrams." Weather and Forecasting 22: 651-661. DOI:10.1175/WAF993.1

90% interval *around the diagonal*, computed using binomial percentiles.

Here showing T+72, 12-member lagged ensemble for all of 2020-2022 using 8 cycles per day
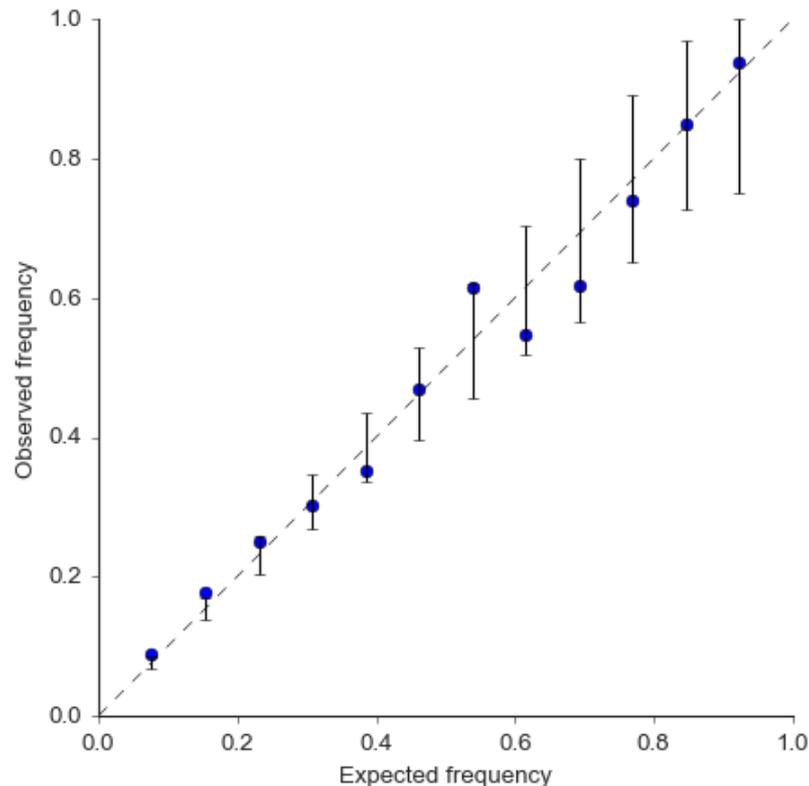
Ensemble consistency diagram

Consistency bars

Ensemble consistency diagram

90% interval around the diagonal, computed using binomial percentiles.

Here showing T+72, 12-member lagged ensemble for *DJF 2023-24 using 4 cycles per day* – i.e. **much smaller sample size**

Summary measures

- The reliability component of the Brier score can be viewed a weighted sum of squared distances from the diagonal of the conventional reliability diagram.

- This also can be misleading for small ensembles.

- Calculating an analogous quantity for the ensemble consistency diagram could give a summary measure ('ensemble miscalibration'?).

- Might this lead to a decomposition of the *fair* Brier score…?

- It appears not, unfortunately 🙁

Summary

- Conventional reliability (calibration) diagrams are misleading for small ensembles

- Including the verifying observation in the conditioning overcomes this, giving an ensemble consistency diagram

- Consistency bars aid interpretation

- Paper in preparation for *Meteorological Applications* special issue on verification (deadline 31 Dec 2024)

roger.harbord@metoffice.gov.uk