Incompatibility of equitability and propriety for the Brier score

Ian T. Jolliffe and David B. Stephenson
Department of Meteorology
University of Reading

1. Definitions – Brier score, propriety and equitability
2. Incompatibility of propriety and equitability
3. A probability model for unskilful ensemble forecasts
4. Which is the best baseline to choose?

Definition of the Brier score

Suppose it is required to give a probability forecast of a binary event – the forecast issued on the i^{th} occasion, $i = 1,2, \ldots,n$, says that there is a probability p_i that the event will occur. Let $x_i = 1$ if the event occurs and $x_i = 0$ if it doesn't. Then the Brier score is given by:

$$BS = \frac{1}{n}\sum_{i=1}^{n}(p_i - x_i)^2$$

The Brier Skill Score

The BS can be converted to a skill score BSS by the linear transformation:

$$BSS = 1 - \frac{BS}{BS_{ref}}$$

where BS_{ref} is the Brier score for some unskilful reference forecast

Definition of *hedging* and *proper* scores

- ‘Hedging’ is when a forecaster gives a forecast different from his/her true belief because he/she believes that the hedged forecasts will improve the score on a measure used to verify the forecasts. Clearly hedging is undesirable.
- A (strictly) proper score is one for which the forecaster (uniquely) maximises the expected score by forecasting his/her true beliefs, so that there is no advantage in hedging.
- BS and BSS are strictly proper.

Definition of *equitability*

- A score is equitable if it takes the same value (often chosen to be zero) for all unskilful forecasts of the type
  - Forecast the same probability all the time or
  - Choose a probability randomly from some distribution on the range [0,1]
- Equitability is desirable – if two sets of forecasts are made randomly, but with different random mechanisms, one should not score better than the other
- The reference forecast used in constructing BSS has a zero value of BSS, by definition

Propriety and equitability are incompatible

- Many possible scores are not proper – BS is one of relatively few that are
- Equitability is even harder to achieve – symmetric scores (those for which the same amount of over- or under-estimation is penalised equally) can only be equitable if the long-run probability of the event, θ (climatology), is 0.5
- It can be shown (new result) that it is impossible to achieve propriety and equitability simultaneously

A probability model for unskilful ensemble forecasts

1. The occurrence of the event is represented by a Bernouilli random variable x, with probability $P(x=1) = \theta$ (climatology)
2. The ensemble with m members is generated from a Binomial distribution with m trials and probability of success ϕ, and the probability forecast is the proportion of successes, p
3. 1 and 2 are independent, so the forecast is unskilful

$$x \sim Be(\theta)$$

$$mp = r \sim Bin(m, \phi)$$

Expected Brier score

The model allows us to calculate the mean (expected) Brier score:

$$E(BS) = E((x-p)^2)$$

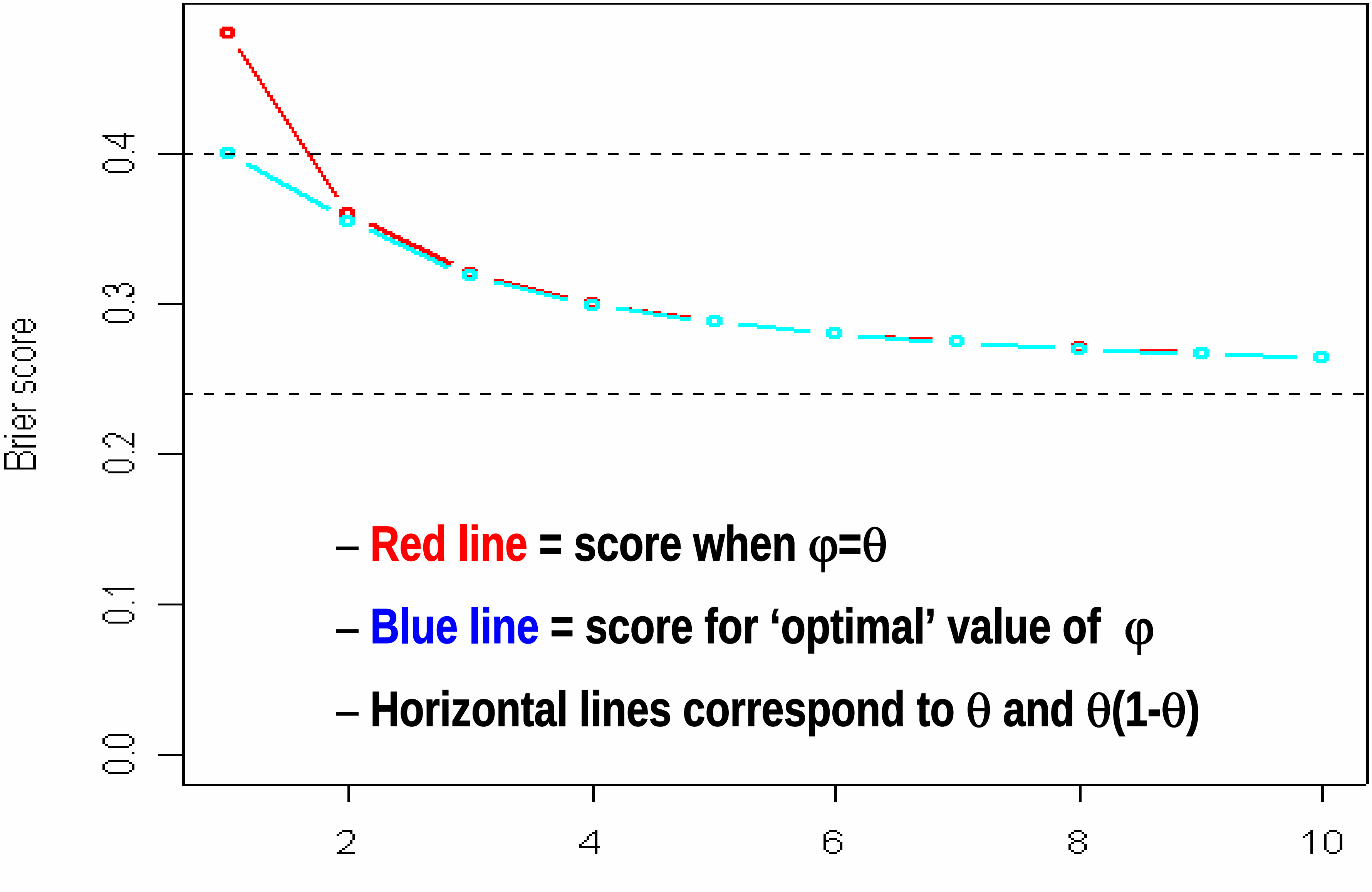
$$= (E(x) - E(p))^2 + \mathrm{var}(x) + \mathrm{var}(p)$$

$$= (\theta - \phi)^2 + \theta(1-\theta) + \frac{\varphi(1-\varphi)}{m}$$

Properties of the mean score

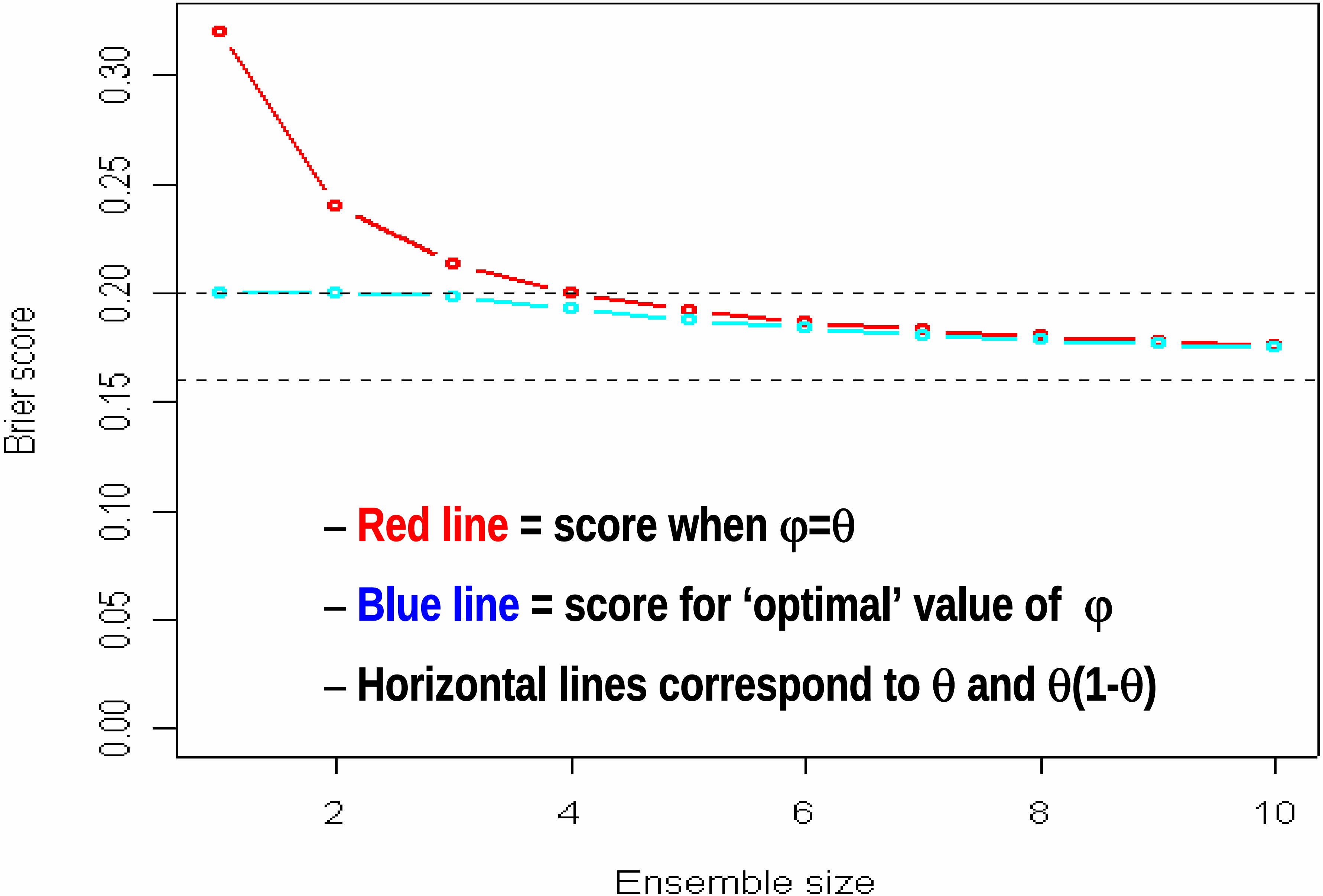
- The smallest mean Brier score is not achieved by *climatology* $\varphi=\theta$ (except when $\theta=0.5$).
- The smallest mean Brier score is obtained for the forecast probability $\varphi=\theta + (2\theta-1)/2(m-1)$ – i.e. φ shifted slightly towards 0 or 1, depending on whether $\theta<0.5$ or $\theta>0.5$.
- If we use this choice as a reference forecasts, then $E(BSS)\leq 0$ for all random forecasts of this type.
- The m=1 special case issues probabilities of 0 and 1 and the Brier score is then equal to one minus the proportion correct. The formula for optimal ϕ breaks down for m=1. Here it is optimal to hedge to 0 or 1 depending on whether $\theta < 0.5$ or $\theta > 0.5$.
- The mean Brier score is the same for $(1-\theta)$ as for θ.

θ=0.4



- **Red line = score when φ=θ**
- **Blue line = score for 'optimal' value of φ**
- **Horizontal lines correspond to θ and θ(1-θ)**

θ=0.2

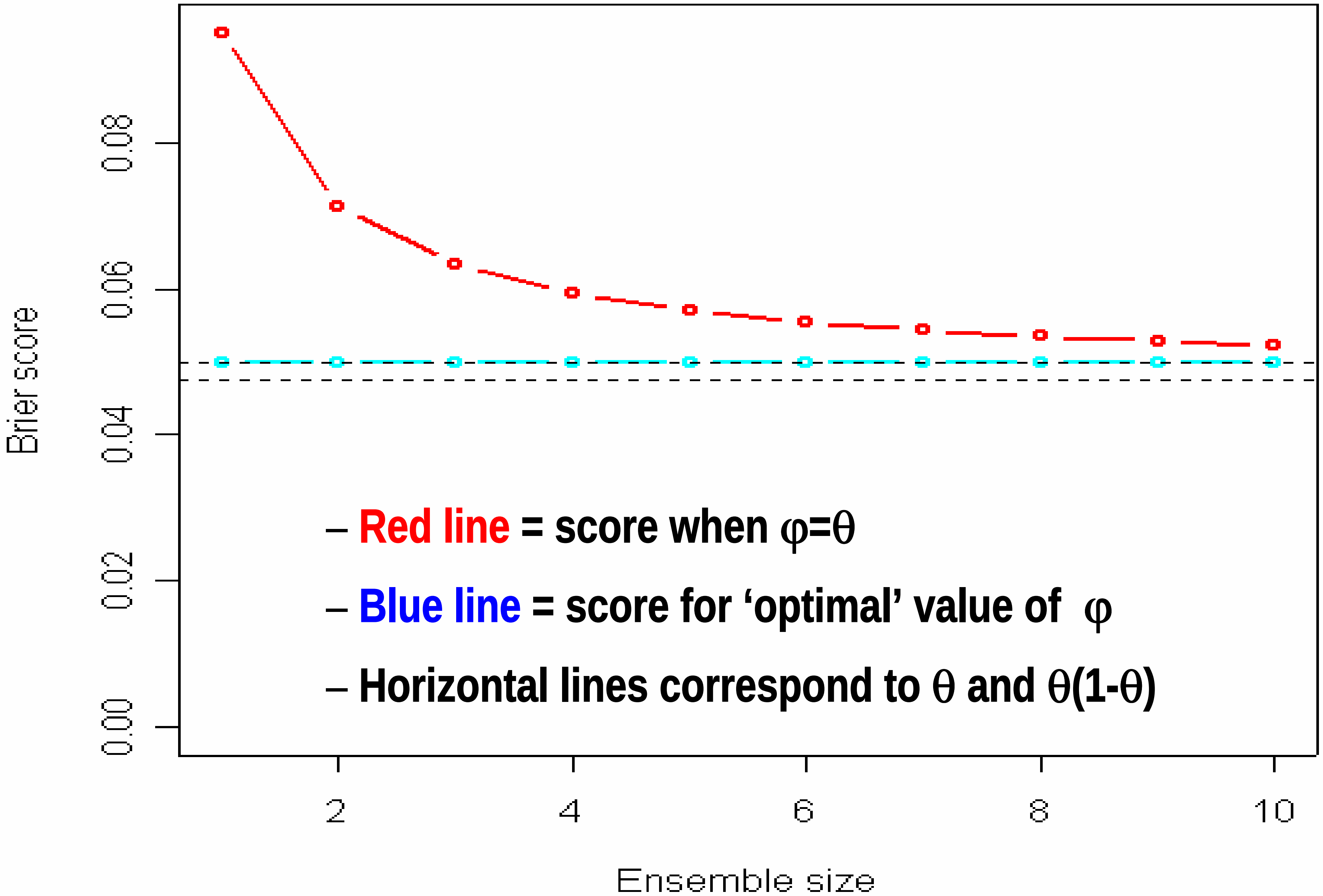


– Red line = score when $\varphi=\theta$

– Blue line = score for 'optimal' value of φ

– Horizontal lines correspond to θ and $\theta(1-\theta)$

$\theta=0.05$


Brier score
0.00
0.02
0.04
0.06
0.08
2
4
6
8
10
Ensemble size
– Red line = score when φ=θ
– Blue line = score for 'optimal' value of φ
– Horizontal lines correspond to θ and θ(1-θ)

Conclusions from plots

- Lowest overall score is when $\phi=\theta$ and $m=\infty$
- All mean scores $\rightarrow \theta(1-\theta)$ as ensemble size $m\rightarrow\infty$
- Greatest 'improvement' compared to climatology occurs when m is small and θ is far from 0.5
- For large ensembles there is little improvement except for extreme events (θ close to 0) or very common events (θ near 1)

Possible reference forecasts

- Minimum score, based on ensemble of size m. $\varphi = \varphi_{min}$. All ensemble-based random forecasts have expected scores ≤ 0.
- Müller et al[†]. – based on ensemble of size m. $\varphi = \theta$. Some ensemble-based forecasts have expected scores > 0.
- Mason[‡] – does not like negative scores for some unskilful forecasts (some forecasts with skill will also have negative values). Chooses a reference forecast so that (most) unskilful forecasts have non-negative values.
- Climatology – Ignore the ensemble and always forecast θ. Traditional. All constant probability forecasts, as well as all ensemble-based random forecasts have expected scores ≤ 0. Equivalent to $\varphi = \theta$; $m=\infty$.

† In press, Monthly Weather Review. ‡ In press, Journal of Climate.

Concluding remarks

1. No proper score is equitable
   → Different unskilled forecasts give different proper scores.
   → It is not clear how best to construct a proper skill score, but if forecasts are based on ensembles our strategy seems sensible.
2. We have assumed that propriety is essential. If it is abandoned, equitability can be achieved.
3. Some of the ideas can be extended to more than two categories via the rank probability score.
4. There are parallel, but somewhat different considerations for deterministic binary forecasts.