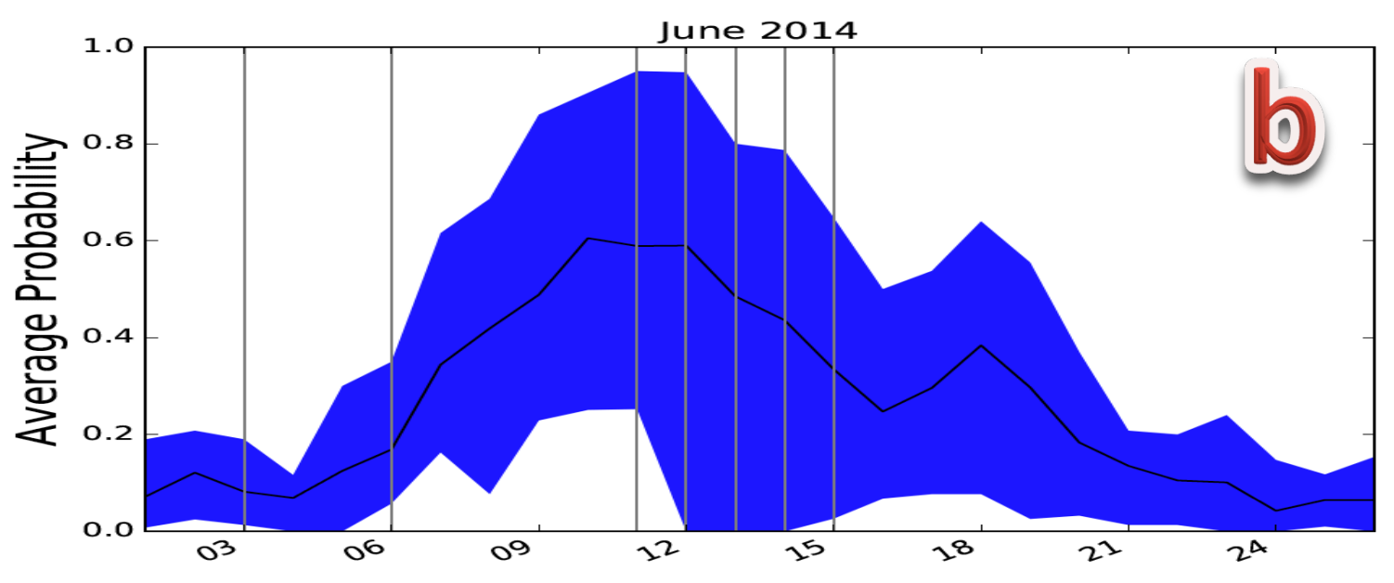
Modeling Ensemble Forecasts of Solar Flares

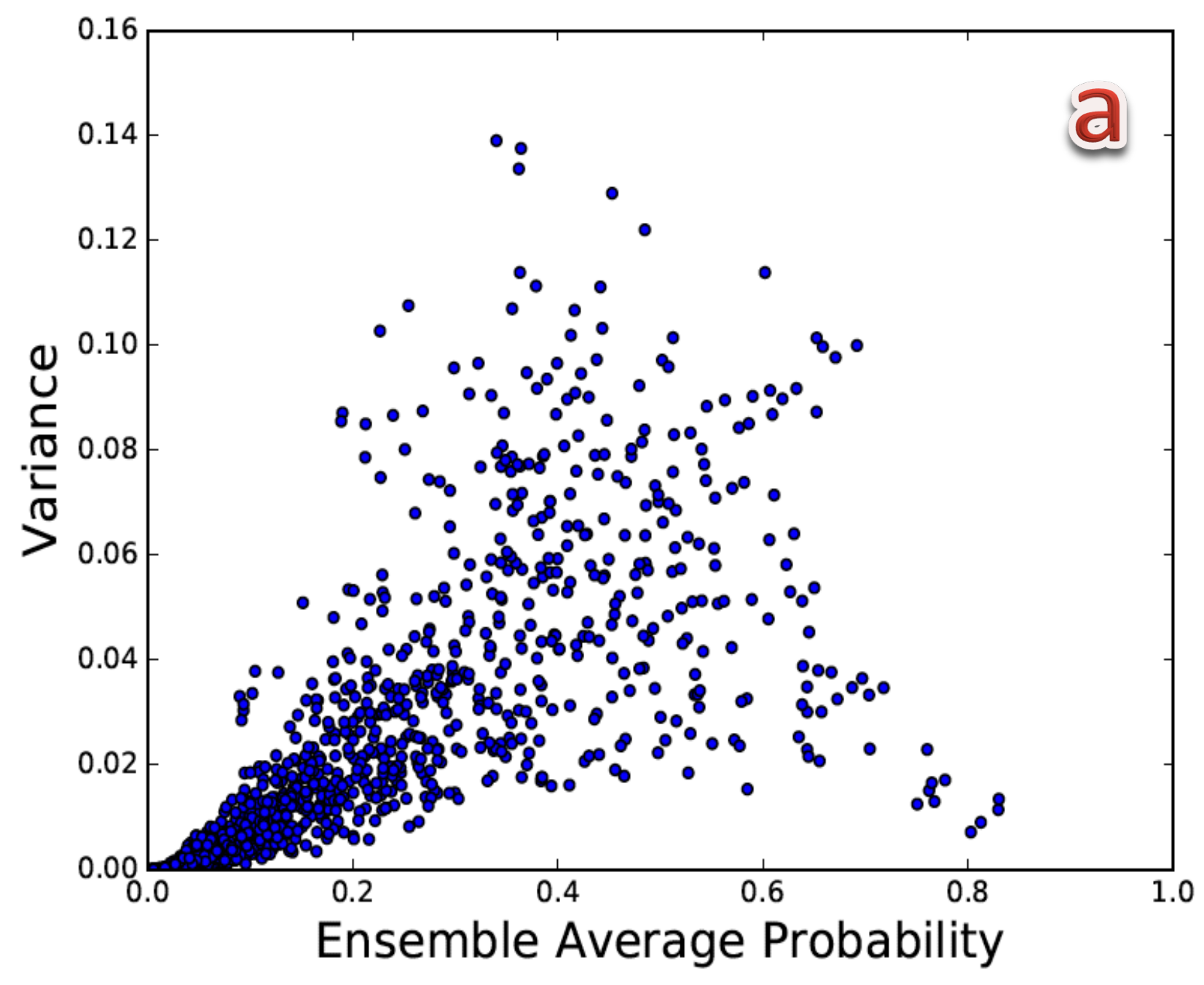
J.A. Guerra[1], S. Murray[1], A. Pulkkinen[2], and D. S. Bloomfield[3]
[1]Trinity College Dublin, [2]NASA/GSFC, [3]Northumbria University

1. Motivation

It is common to observe that for given photospheric conditions − somehow favorable to the production of flares − the likelihood of observing a major flare can vary (sometimes significantly) from method to method. Figure 1 shows that the spread of values around the mean value is only small for very low or very high probabilities. For values in between, the probability of observing a flare can have a wide range of values. In such cases, a combination of such values will be closer to the real probability. A simple average often proves to be superior to any individual value. However, in most cases, the average is not the best performing combination. In this poster we investigate the construction of several ensemble predictions in order to provide some guidance in an operational environment about the best performing combination according to specific needs of any end user.

Figure 1: (a) Variance of the ensemble members probabilities as a function of their average; (b) time series of ensemble average probability. For low (P < 0.1) and high (P > 0.8) probabilities, most ensemble seem cluster around the average. For probabilities in between (0.1< P <0.8), members values can be spread widely around the average (see Figure 1.b)





3. Ensemble Construction

Probabilistic forecasts from the ensemble members *P* = {P_{MAG4}, P_{ASSA}, P_{ASAP}, P_{NOAA}, P_{MOSWOC}, P_{FPS}} are linearly combined as

$$P^c(w;t) = \sum_i w_i P_i(t)$$

with *w* = {w_{MAG4},w_{ASSA},w_{ASAP},w_{NOAA},w_{MOSWOC},w_{FPS}}. Combination weights are constrained to

$$\sum_i w_i = 1$$

Therefore, problem is reduced to determined w by optimization of a performance metrics. Table 2 shows the list of performance metrics employed.

Table 2: Performance metrics used for the optimization during the ensemble construction. All metrics are well know and widely used for validation of forecasts. Categorical metrics are calculated via the 2x2 contingency table after applying a decision threshold to probabilistic forecasts.

Probabilistic	Categorical
Brier Score	Brier Score
Linear Correlation Coeff. (r)	True Skill Score (TSS)
Mean Absolute Error (MAE)	Heidke Skill Score (HSS)
ROC Score	Perc. Correct (PC)
	Prob. of Detection (POD)
	Prob. of False Detection (POFD)
	False Alarm Rate (FAR)

4. Concluding Remarks

Table 3 list the ensemble predictions according to the overall performance across the four metrics: ROC area (Figure 4), Reliability and Resolution (Figure 5), and Brier score.

Ensemble Forecast	ROC area	Brier score	Reliability	Resolution	Uncertainty
	(0 is worse)	(0 is better)	(0 is better)	(0 is worse)	
Brier	0.855844	0.1076	0.001996	0.0365	0.1421
ROC score	0.8571562	0.1097	0.003513	0.03593	0.1421
TSS	0.8572	0.1097	0.0035	0.0359	0.1421
PC	0.8420	0.1141	0.0076	0.0356	0.1421
Brier (Categorical)	0.8420	0.1141	0.0076	0.0356	0.1421
HSS	0.8379	0.1141	0.0081	0.0361	0.1421
Average	0.8237241	0.1165	0.002729	0.02838	0.1421
POFD	0.8339	0.1152	0.0047	0.0317	0.1421
Correl	0.8098726	0.1183	0.004479	0.02832	0.1421
MAE	0.769607	0.1259	0.006506	0.0227	0.1421
POD	0.8108	0.1358	0.0213	0.0276	0.1421
FAR	0.5695	0.1511	0.0169	0.0080	0.1421

Table 3: Rank of the ensemble forecast according to their overall performance.

- Top 3 performing ensemble (highlighted) are obtained by optimizing the Brier score, ROC score (ROC curve area), and the True Skill Score.
- In this study at least 7 ensembles performed better than the average ensemble.
- 9/11 ensemble forecasts included at least one of the two human-influenced methods (NOAA & MOSWOC). Most common automated method in the ensembles is MAG4 (5/11).
- Results can be sensitive to gaps in the data (e.g. ASAP and FPS)

Future Work:

- Include more probabilistic metrics: Ranked Probability Score, Reliability, Resolution.
- Test the sensibility of the results to precision of combination weights and thresholds.
- Determine the influence of forecasts cadence in the resulting ensemble.

Contact information: jordan.a.guerra@tcd.ie / jordan.guerra@gmail.com

2. Forecasting Methods and Data

The forecasting methods included in the ensembles are listed in Table 1. Full-disk probabilistic forecasts for the occurrence of a M-class flares between 2013 and 2016 are used. Figure 1 displays time series (a) and histograms (b) for each method.

Method	Responsible	Predictor	Notes
MAG4	U. Of Alabama	WL_{SG}	Automated
ASSA	Korean Space Weather Center	McIntosh Class	Automated
ASAP	U. Of Bradford, UK	McIntosh Class	Automated
NOAA	SPWC NOAA, US	McIntosh Class	Human influenced
MOSWOC	Met Office, UK	McIntosh Class	Human influenced
FPS	TCD - SolarMonitor	McIntosh Class	Automated

Table 1: Flare forecasting methods members of the ensemble calculations. NOAA, MOSWOC, and FPS produce forecast at 24 hours cadence while MAG4, ASSA, and ASAP produce forecasts at least hourly. Data for MAG4, ASSA, ASAP, and NOAA was obtained from iswa.ccmc.gsfc.nasa.gov. MOSWOC forecasts were provided by the Met Office and FPS can be obtained from solarmonitor.org.

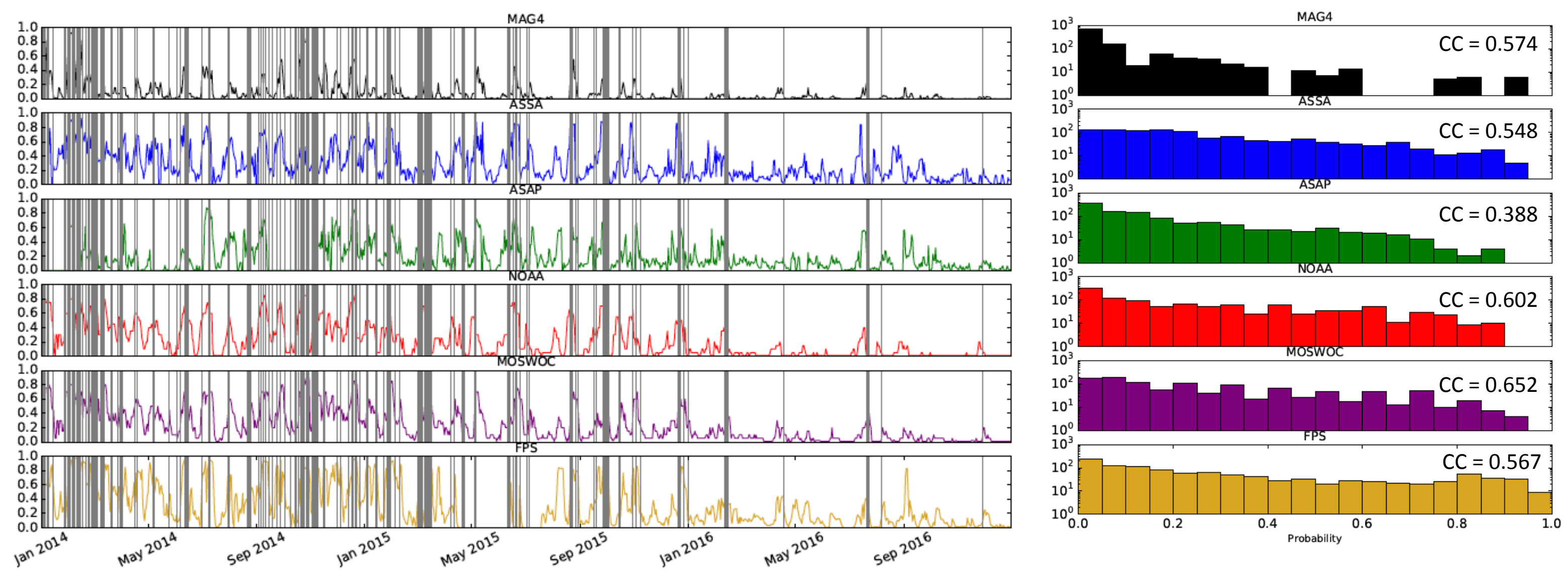


Figure 2: (a) Time series of probabilistic forecasts for each method. Vertical grey lines correspond to the occurrence of flares during the study time interval. All probabilities correspond to the likelihood of observing a M-class flare within the next 24 hours. (b) Distribution of probability values for each method. All methods show weak to moderate levels (CC = 0.3 − 0.7) of correlation among all the ensemble members.

4. Preliminary Results

Figure 3 shows the determined combination weights for probabilistic metrics (a) and categorical metrics (b).

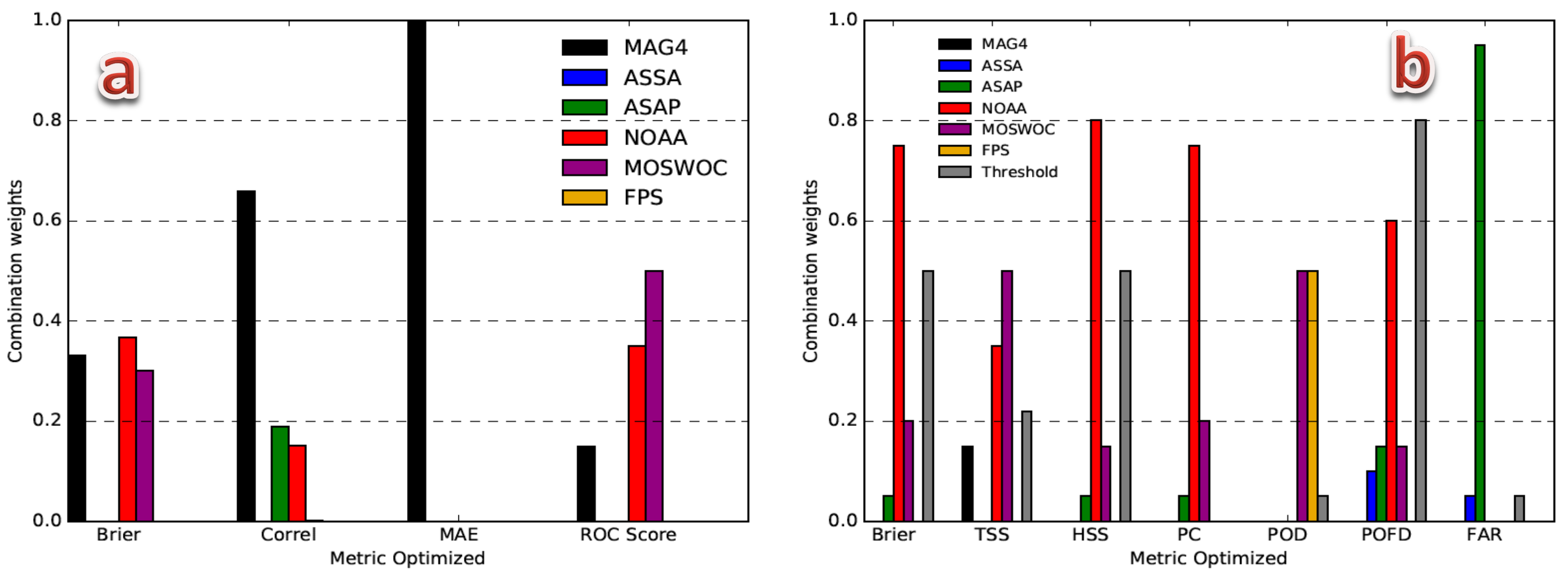


Figure 3: (a) Combination weights obtained using probabilistic metrics. In this case, each ensemble can be constructed using up to three members methods. Methods included in a particular ensemble and their corresponding combination weights are completely chosen so the corresponding metric is optimal. (b) Combination weights obtained using categorical metrics. For the categorical case, curves of optimized metrics as a function of threshold probability are calculated. The weights displayed in Figure 3(b) correspond to the threshold value (grey bar) that produced the overall optimal metric value. For categorical metrics a minimum of 2 and a maximum of 4 methods were included in the ensembles.

The performance of the constructed ensembles is determined by the ROC curve (Figure 4) and Reliability plot (Figure 5).

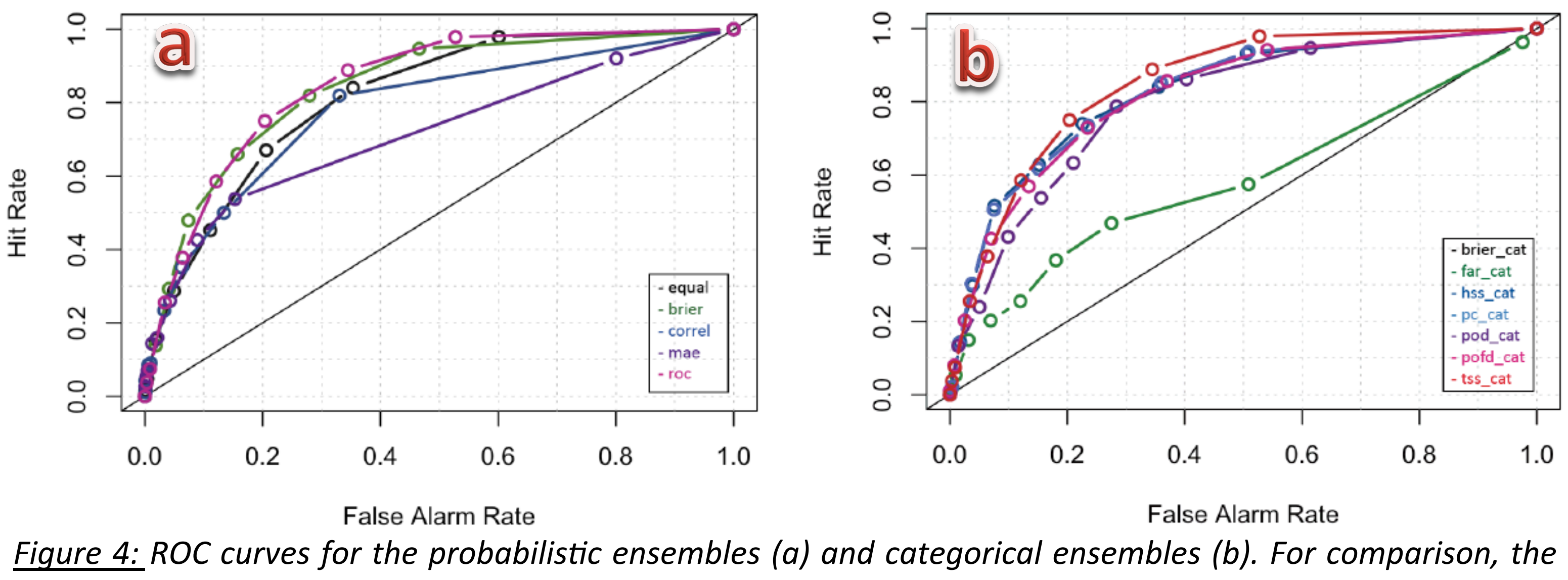


Figure 4: ROC curves for the probabilistic ensembles (a) and categorical ensembles (b). For comparison, the average ensemble forecast is included in Figure 4(a). For the probabilistic case, it's expected that ROC ensemble displays the best ROC curve. However, Brier the ensemble score show a very similar ROC curve as the ROC ensemble. For categorical ensembles, optimizing TSS seem to produce the best ROC curve.

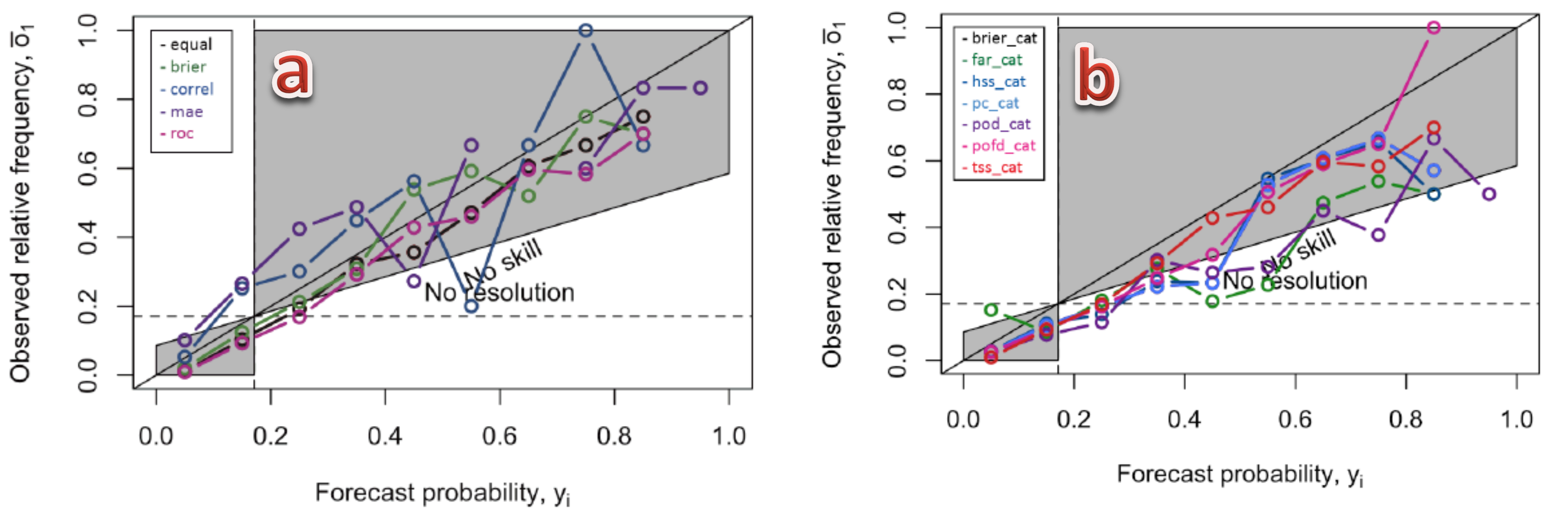


Figure 5: Reliability plots for probabilistic (a) and categorical (b) ensembles. The ensemble average forecast is included in Figure 5(a) for comparison. In both panels, it is observed that all ensemble follow the diagonal in the reliability plot. For the probabilistic case, the Brier, Average, and ROC ensembles appear to follow the diagonal closer than the Correlation and MAE ensembles. For the categorical case, TSS and Brier ensembles show the best Reliability curves.