Verification of predictions of CME arrival time at L1

Ed Pope, Met Office, UK

Verifying predictions of CME arrival time at L1

- Compared **MOSWOC archived forecasts** & CME **Scoreboard average** of methods with the **Scoreboard observed** time
- Data: April-December 2014
- Method:
  - Compare MOSWOC arrival time prediction with observed arrival time on Scoreboard.
  - Produce a MOSWOC contingency table (hit, miss, false alarm, correct rejections).
  - Do same for Scoreboard average.
  - Calculate scores & confidence intervals (CIs) for both approaches.
- Confidence interval: a way of quantifying variation in statistical calculations. If CIs overlap, then you can say that no difference exists between the overlapping X & Y. If CIs don't overlap then you can say, e.g. with 95% confidence X is more skilled than Y.

CME: 2016-04-10T11:00:00-CME-001

Actual Shock Arrival Time: 2016-04-14T06:50Z

Observed Geomagnetic Storm Parameters:
Max Kp: 5.0
CME Note: CME associated with large filament eruption situated close to N18E29 starting around 10UTC.

Predicted Shock Arrival Time	Difference (hrs)	Confidence (%)	Submitted On	Lead Time (hrs)	Predicted Geomagnetic Storm Parameter(s)	Method	Submitted By	
2016-04-14T00:00Z (-7.0h, +7.0h)	-6.83	----	2016-04-11T00:54Z	77.93	----	WSA-ENLIL + Cone (GSFC SWRC)	Yaireska Collado (GSFC)	Detail
2016-04-13T14:00Z	-16.83	----	2016-04-11T05:07Z	73.72	Max Kp Range: -- - 5.0	WSA-ENLIL + Cone (NOAA/SWPC)	Leila Mays (GSFC)	Detail
2016-04-13T18:00Z (-12.0h, +6.0h)	-12.83	30.0	2016-04-11T05:45Z	73.08	Max Kp Range: 4.0 - 6.0	WSA-ENLIL + Cone (Met Office)	Met Office (Met Office)	Detail
2016-04-14T12:00Z (-12.0h, +12.0h)	5.17	----	2016-04-11T12:30Z	66.33	----	Other (SIDC)	Leila Mays (GSFC)	Detail
2016-04-13T04:51Z	-25.98	100.0	2016-04-12T20:30Z	34.33	----	SPM2	Xinhua Zhao (NSSC CAS)	Detail
2016-04-13T12:44Z	-18.10	----	2016-04-12T20:33Z	34.28	----	SPM	Xinhua Zhao (NSSC CAS)	Detail
2016-04-13T18:15Z	-12.58	65.0	---	---	Max Kp Range: 4.0 - 5.5	Average of all Methods	Auto Generated (CCMC)	Detail

Contingency Table

		Observed		
		yes	no	Total
Forecast	yes	***hits***	***false alarms***	***forecast yes***
	no	***misses***	***correct negatives***	***forecast no***
Total		***observed yes***	***observed no***	***total***

Results:
scores used to compare MOSWOC &
CCMC Scoreboard average,
for CME arrival time

Score	MOSWOC (M)	5% CL	95% CL	Score-board average (S)	5% CL	95% CL	A measure of...
Hits	**33**			**27**			Number of times a yes forecast was a yes occurrence.
Misses	**9**			**0**			Number of times a no forecast was a yes occurence.
False alarms	**6**			**12**			Number of times a yes forecast was a no occurence.
Correct rejections	**7**			**9**			Number of times a no forecast was a no occurrence.
Hit rate (probability of detection- POD)	**0.79**	0.68	0.88	**1**	1	1	Discrimination What fraction of observed yes events were correctly forecasted? = hits/(hits + misses) 1=perfect. Sensitive to hits. Ignores false alarms. Good for rare events. Use with FAR. *S=perfect. Ranges don't overlap.*
False alarm rate (Probability of False Detection- POFD)	**0.46**	0.23	0.7	**0.57**	0.4	0.75	Discrimination What fraction of the observed no events were incorrectly forecasted as yes? Conditioned on observations not forecasts. 0=perfect. Sensitive to false alarms. Ignores misses. *M better than S, however ranges overlap.*
False alarm ratio (FAR)	**0.15**	0.07	0.25	**0.31**	0.19	0.43	Reliability What fraction of the predicted yes events didn't occur? = false alarm/(hits + false alarms) 0=perfect. Sensitive to false alarms. Ignores misses. Use with POD. M *is better than S. Ranges just overlap.*
Proportion correct	**0.73**	0.64	0.82	**0.75**	0.65	0.83	Accuracy Correct predictions of both events & non-events. = (hits + correct negatives)/total forecasts Possible to obtain a higher PC by not forecasting rare events at all. *Comparable for both.*
Base rate	**0.76**	0.67	0.86	**0.56**	0.46	0.69	Event frequency/sample climatology. The uncertainty in the occurrence of the observations. = observed yes's/total
Forecast rate	**0.71**	0.6	0.8	**0.8**	0.73	0.9	
Threat score	**0.69**	0.58	0.79	**0.69**	0.57	0.81	Accuracy How well did the forecasted yes events correspond to the observed yes events? 0=no skill, 1=perfect. Sensitive to hits, penalises misses & false alarms. 0.69 means that more than half of the events were correctly forecasted. *Comparable for both. Ranges overlap.*
Bias score	**0.93**	0.79	1.09	**1.44**	1.24	1.76	Bias How did the forecast frequency of yes events compare to the observed frequency of yes events? 1=perfect. Measures the ratio of the frequency of forecast events to frequency of observed events. Doesn't measure how well forecast corresponds to observations (only measures relative frequencies). *M<1 so under-forecasting. S>1 so over-forecasting. Ranges don't overlap.*
Equitable threat score	**0.18**	0.04	0.34	**0.3**	0.16	0.47	Skill How well did the forecast yes events correspond to the observed yes events (accounting for hits due to chance in the threat score)? 0=no skill, 1=perfect. Sensitive to hits. Two approaches are comparable & ranges overlap.
Heidke score	**0.3**	0.07	0.51	**0.46**	0.27	0.64	Skill What was the accuracy of the forecast relative to that of random chance? Range -1 to 1. 0= no skill. 1=perfect. *Suggests some skill in both forecasting approaches. M slightly lower than S, however ranges ovelap.*
Peirce score	**0.32**	0.08	0.57	**0.43**	0.25	0.6	Skill How well did the forecast separate the yes events from the no events? Similar to Heidke. Range -1 to 1. 0=no skill. 1=perfect. Peirce may be more useful for more frequent events. *The two approaches are comparable & ranges overlap.*

Contingency Table

		Observed yes	Observed no	Total
Forecast	yes	*hits*	*false alarms*	*forecast yes*
	no	*misses*	*correct negatives*	*forecast no*
Total		*observed yes*	*observed no*	*total*

Summary

- Only a short period of data analysed – rerun with more data, preferably several years
  - may help to reduce confidence intervals
  - as indication of whether skill has changed over time (improved through experience/ got worse through losing STEREO?)

- Difficult to strongly distinguish differences between MOSWOC & Scoreboard average.

- Suggestion that NASA are over-predicting (high hit-rate & high false alarm rate) & MOSWOC are under-predicting.

- Ambiguity of ‘hit’, e.g. when CMEs in quick succession.

- Would be interesting to do cost-benefit analysis, since false alarms are potentially expensive for users.

- Verification definitions: http://www.cawcr.gov.au/projects/verification/