Skill in real-time solar wind shock forecasts

J. B. Mozer

Space Weather Center of Excellence, Space Vehicles Directorate, Air Force Research Laboratory, Sunspot, New Mexico, USA

W. M. Briggs

Weill Medical College, Cornell University, New York, New York, USA

Received 3 January 2003; revised 21 March 2003; accepted 3 April 2003; published 27 June 2003.

[1] Forecasts of 96 shocks in the solar wind at the L1 point made by the HAF kinematic solar wind model made between 5 February and 31 December 2001 are compared with algorithmically determined shocks from real-time data from the SWEPAM and MAG instruments aboard the ACE spacecraft. Traditional measures of forecast skill used by the meteorological community are applied to these forecasts and observations and indicate modest skill in the HAF model. Details of forecast skill are determined using the Briggs-Ruppert skill score which incorporates the costs of forecast error. This new metric is shown by the present example to be a potentially valuable tool to the customer of space weather forecasts in that it focuses on the actual application of the forecasts and identifies economic regimes where a given forecast is potentially valuable. *INDEX TERMS:* 2139 Interplanetary Physics: Interplanetary shocks; 2164 Interplanetary Physics: Solar wind plasma; 2194 Interplanetary Physics: Instruments and techniques; 2102 Interplanetary Physics: Corotating streams; 2109 Interplanetary Physics: Discontinuities; *KEYWORDS:* solar wind shock forecasts, Hakamada-Akasofu-Fry solar wind model, forecast skill

Citation: Mozer, J. B., and W. M. Briggs, Skill in real-time solar wind shock forecasts, *J. Geophys. Res.*, *108*(A6), 1262, doi:10.1029/2003JA009827, 2003.

1. Introduction

[2] A primary goal of space weather research is to accurately predict the onset and magnitude of geomagnetic storms with enough lead time, accuracy, and reliability so that decision makers can use such predictions to minimize the harmful effects of such storms on, for example, spacecraft, communication, and navigation systems. A fundamental component of any space weather forecast capability is the ability to diagnose and predict the solar wind and the structures that lie therein. Shocks in the solar wind plasma and deformations of the ambient magnetic field associated with Coronal Mass Ejections (CMEs), stream-stream interactions, and Corotating Interaction Regions (CIRs) herald the arrival of geomagnetic storms [c.f., *Luhmann*, 1997 and references therein]. The ability to forecast these events is crucial to a successful space weather capability.

[3] Several models aimed at forecasting solar wind conditions at 1 AU have been developed and show varying degrees of skill. The basis of these models varies substantially. For example, the Shock Time Of Arrival (STOA) model [*Dryer and Smart*, 1984; *Smart and Shea*, 1984, 1985] is based on a similarity theory of blast waves and seeks to predict only a single quantity (shock arrival time). Also used for this purpose is the Interplanetary Shock Propagation Model (ISPM) which is a statistical and parametric model based on 2.5-dimension magnetohydrodynamic (MHD) simulations [*Smith and Dryer*, 1990]. Slightly more complex is the kinematic solar wind model of Hakamada-Akasofu-Fry (HAF) that in addition to shock arrival time, provides the vector direction and magnitude of shocks as well as forecasts of solar wind speed, density, dynamic pressure and interplanetary magnetic field [*Fry et al.*, 2001, and reference therein]. Even further down this spectrum of complexity are the fully three-dimensional (3-D) MHD codes that are currently under development and promise to provide even greater accuracy and fidelity in solar wind forecasting [c.f., *Riley et al.*, 2001; *Gombosi et al.*, 2001].

[4] All of these solar wind models require similar inputs in order to produce a forecast. They all start with an observed event, such as an optical or X-ray flare, or a white-light coronagraph observation of a CME. These observations serve as indicators that a CME or some other disturbance may have been launched from the Sun into the interplanetary solar wind field. Metric Type II radio bursts, which are considered to be the signatures of shocks traveling outward through the corona are used to determine the initial shock speed and duration. The location of the initial flare on the Sun provides information on the direction of the disturbance. In addition to these data the STOA model also requires ambient solar wind velocity, which is available from Sun-orbiting satellites and HAF requires the magnetic field on the so-called potential field source surface, which is derived from magnetogram observations of the Sun.

0148-0227/03/2003JA009827$09.00

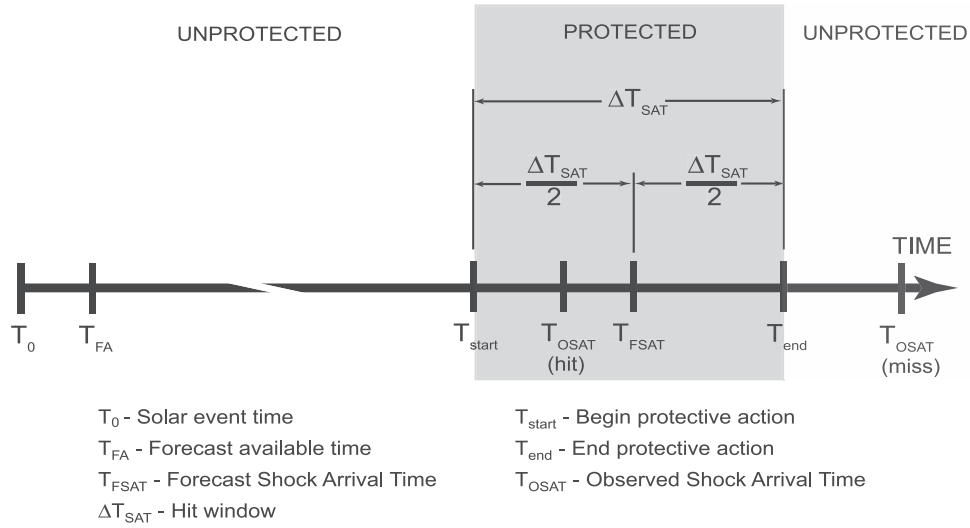

Figure 1. Hypothetical shock arrival forecast timeline.

[5] Unfortunately, very few observational data are available on the propagation and evolution of structures in the solar wind as they travel between the Sun and Earth. Currently, Interplanetary Scintillation (IPS) measurements made by looking at the propagation of radio waves from distant sources through structures in the solar wind are the only constant source of such data. In the very near future, however, the Solar Mass Ejection Imager (SMEI) has the potential to greatly improve the quantity of this type of data through the use of polar-orbiting, white-light cameras which will detect CMEs and other dense structures in the solar wind out to and beyond 1 AU [*Webb et al.*, 2002].

[6] Previous studies of solar wind forecast skill have been performed [*Smith et al.*, 2000; *Fry et al.*, 2001; *Sun et al.*, 2002; *Thomson*, 2000]. The present study differs from these in that it focuses only on the HAF model forecasts made in the year 2001, includes forecasted and observed shocks associated with sector transitions and other corotating interaction regions (CIRs), and applies a different set of skill scores to the analysis, with particular emphasis on those that determine forecast value from the user's perspective. These skill scores are discussed in section 2. Details of the data used, including the HAF model forecasts and shock observations are given in section 3. The analysis and conclusions are given in sections 4 and 5, respectively.

2. Forecast Skill Measures

[7] Although traditional measures of forecast skill used by the meteorological community are applicable to solar wind shock forecasting, there are aspects of the problem that distinguish it from conventional weather forecasting applications. Furthermore, since the quality and quantity of space weather forecasts have traditionally been marginal (at best) for many applications where critical decisions must be made, a brief discussion of the skill of space weather forecasts from the point of view of the end-user is warranted here.

[8] Consider the hypothetical example of a satellite operator that has the power to shut down and stow an on-orbit instrument that is sensitive to solar wind shocks or their associated energetic particles. With reference to Figure 1, assume that at some time, T_0, an X-ray flare, and a coincident optical flare are observed in the western hemisphere of the Sun and an initial velocity is determined through an analysis of a Type II radio burst that is also observed near the same time. These events trigger a special run of a solar wind forecast model. At a somewhat later time, T_{FA}, the results of this forecast are made available to the user community, including the satellite operator. The forecast calls for a shock to arrive in the vicinity of Earth at a future time, T_{FSAT}. On the basis of an analysis of the past performance of the forecast model (such as provided in the present work), a shock arrival time window, ΔT_{SAT} is also provided to the end users.

[9] Now, at some time during the period between T_{FA} and the beginning of the shock arrival window, $T_{FSAT} - \Delta T_{SAT}/2$, the user must make a decision to either stow the instrument for protection or to keep it online despite the forecast. One unique aspect of the shock arrival time problem is that there may be no further updates of the forecast in the intervening time, since no additional observations of the shock may be available as it propagates through interplanetary space. (In the future this condition may be improved through the assimilation of data from, e.g., the Solar Mass Ejection Imager, STEREO, or Interplanetary Scintillation measurements.) Assume that the user weighs his options and decides to stow the instrument at time $T_{start} = T_{FSAT} - \Delta T_{SAT}/2$. Furthermore, he programs the instrument to come back online at the end of the time window, given by $T_{end} = T_{FSAT} + \Delta T_{SAT}/2$.

[10] If a shock arrives at Earth during the period forecasted shock arrival time window, then the forecast is a "hit" and a potential catastrophe is averted. If, however, no shock is observed, the forecast is a "miss" (specifically, a false alarm) and the satellite operator suffers a loss due the unnecessary downtime of the instrument. If the expected shock arrives at some time outside of the window, the forecast also fails and the operator may again suffer a loss due to instrument damage.

[11] From this point of view it is clear that for a space weather forecast to be of value, several things must hold: a high rate of forecast hits, a low rate of misses (both false alarms and unforecasted events), a small hit window, and

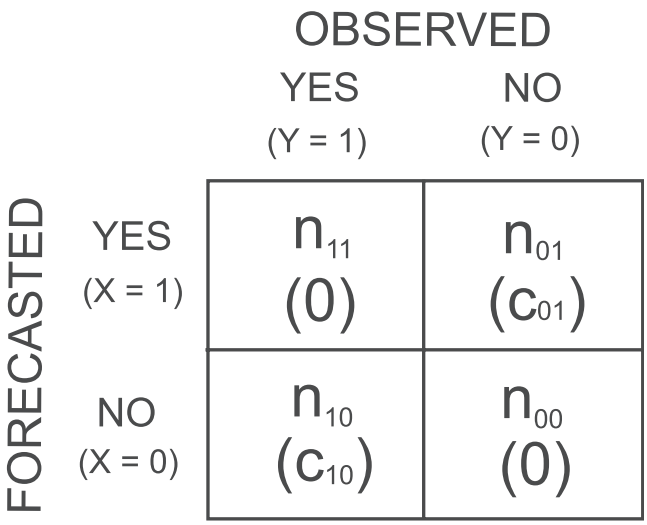

Figure 2. 2 × 2 contingency table and associated loss table (denoted by c_{ij}).

sufficient lead time of forecasts. It is the purpose of the present work to evaluate the HAF solar wind model in terms of all of these criteria, keeping in mind that the costs of each criteria may be different for each end user of the forecast.

[12] Meteorologists have gauged the utility of various weather forecasts using a wide variety of metrics, or skill scores, which relate predicted quantities to those actually observed [c.f., *Wilks*, 1995]. The particular skill score used to evaluate a forecast should always be chosen to best represent the actual use of a forecast. For example, dichotomous, or ''yes/no'' forecasts of a discrete predictand, which is the case for forecasts of solar wind shock arrival (the shock will or will not occur in a given time window), lend themselves to a description in the form of a 2 × 2 contingency table (see Figure 2).

[13] A contingency table relates events that are predicted to those observed. The goal of any forecast system is to maximize the number of forecast ''hits'' (the prediction agrees with observation, whether the event occurred or not) and minimize the number of ''misses'' (which can be either a predicted event that was not observed or an observed event which was not predicted). A complete description of commonly used skill scores is given by *Wilks* [1995] and only a brief outline is given here.

[14] The simplest skill score that can be derived from a 2 × 2 contingency table is the hit rate which gives the proportion of correct forecasts,

$$H = \frac{n_{11} + n_{00}}{n}, \tag{1}$$

where $n\Sigma_{ij} = n_{ij}$. This metric is useful when correct positive and negative forecasts hits are of equal value. A hit rate of 1 represents a perfect forecasting system and the worst possible hit rate is 0. A frequently used alternative to hit rate is the so-called threat score,

$$TS = \frac{n_{11}}{n_{11} + n_{01} + n_{10}}, \tag{2}$$

which is particularly useful in cases where the occurrence of events is much less frequent than nonoccurrence (which is the case in solar wind shocks). The interpretation of threat score is that it expresses the ratio of events correctly forecasted to those that were either forecasted or observed. In this way, correct null forecasts are removed. Again, the best threat score is 1 and the worst is 0.

[15] Similarly, the ratio of the number of occasions that an event occurs to the number of times that it was forecast is captured by the Probability of Detection (POD),

$$POD = \frac{n_{11}}{n_{11} + n_{01}}. \tag{3}$$

As with the H, and TS, a perfect forecast is represented by POD of unity.

[16] The effect of forecasted events that are not observed is quantified by the false-alarm rate,

$$FAR = \frac{n_{01}}{(n_{01} + n_{11})}. \tag{4}$$

In this case, a $FAR = 0$ is the best possible forecast and $FAR = 1$ is the worst.

[17] A global measure of a forecast system is given by the bias,

$$B = \frac{n_{11} + n_{01}}{n_{11} + n_{10}}, \tag{5}$$

which simply is the ratio of the total number of yes forecasts to yes observations, whether or not the observations and forecasts agree. Bias is used to determine if a forecast system is consistently over or under forecasting events. A bias value of 1 is the ideal.

[18] The standard skill scores represented by equations (1)–(4) are useful as broad indicators of the performance of a particular forecast system; however, the value of a hit or the severity of a miss in a forecast depends heavily on how the forecast is being used. For example, as discussed in the hypothetical example above, if the action taken by a satellite operator is to shut down systems when an event is forecasted, he may be very sensitive to false alarms due to unnecessary losses in satellite uptime. On the other hand, a large scale power blackout caused by ground-induced currents associated with an unforecasted event may be more critical to long-haul power line operators.

[19] A novel forecast skill score that incorporates the cost to the user of incorrect forecasts has been proposed by W. M. Briggs and D. Ruppert (personal communication, 2002, hereafter referred to as BR). This work generalizes that of *Thomson* [2000] and is outlined in Appendix A, who first applied decision analytic techniques to space weather forecasts. Here, we extend the work of Thomson by allowing the forecast user to calculate and to assess the statistical significance of a simple economic skill score. We also introduce new graphical methods that can be used in assessing forecast skill (see *Schervish* [1989] and *Wilks* [2001] for examples of graphical skill assessment).

[20] The BR skill score is given by

$$K_\theta = \frac{n_{11}(1-\theta) - n_{01}\theta}{(n_{11} + n_{10})(1-\theta)}, \tag{6}$$

where θ is a measure of loss given by

$$\theta = \frac{c_{01}}{c_{01} + c_{10}}, \tag{7}$$

where c_{01} is the cost of a false positive forecast, and c_{10} is the cost of a false negative forecast. These costs are dependent on the assessment made by the user of the forecast. Since different decision makers may assign different values to these quantities, the utility of a forecast, as measured by this skill score, may be different for different users. Details of the use of this skill score are given by W. M. Briggs and D. Ruppert (personal communication, 2002).

3. Data

3.1. Forecasted Events

[21] The HAF model produces forecasts of shocks and their arrival time at the so-called Lagrangian 1, or L1 point located roughly .01 AU sunward of the Earth. In recent years the HAF has been run regularly as part of National Oceanic and Atmospheric Administration/Space Environment Center (NOAA/SEC) unofficial "Fearless Forecast." The model is run hourly to produce nominal forecasts of ambient solar wind based on input boundary conditions (e.g., the potential field source surface) up to forty days in the future [c.f., *Sun et al.*, 2002]. When a significant solar event is observed, it is analyzed by researchers at NOAA/SEC in order to determine inputs to the HAF model, such as initial shock velocity, as described in section 1. These inputs subsequently form the basis of forecasts of event-driven transients in the solar wind.

[22] The basic output of HAF consists of solar wind velocity and density (from which a dynamic ram pressure can be computed) and interplanetary magnetic field components. A Shock Searching Index [*Fry et al.*, 2001] given by

$$SSI_H = \log \frac{\Delta P}{P_{min}} \tag{8}$$

where P is either the dynamic pressure or the momentum flux, ΔP is a change in P, and P_{min} is the local minimum of P over a window of time steps. Shocks are identified when SSI_H exceeds a given threshold, which is determined empirically. For a given shock, the Shock Arrival Time (SAT) is simply the time of the first occurrence of SSI_H greater than the specified threshold. For the present study, a nominal threshold of -0.35 was chosen based on earlier studies [*Fry et al.*, 2003].

[23] Because the HAF model is run hourly and shocks take roughly 3–5 days to propagate from the Sun to the Earth, multiple forecasts of the same shock are frequently produced by the system. The potential field source surface that is used to provide the initial magnetic field boundary conditions to the model can change over a period of days as new observations are available. This leads to a potential discontinuity in the model output which also must be accounted for. To resolve the ambiguity in SAT due to these two effects, we have chosen to select the shock forecast that was made first after the solar events were identified. From the perspective of a user of a space weather forecast, this is equivalent to studying the forecast skill of the longest lead-time forecast that can be produced.

[24] The list of the shocks forecasted by the HAF model and distributed in real-time via e-mail to interested parties in 2001, with the longest lead time and $SSI > -0.35$ is given in Table 1. Also shown in the table are the time (T_0) and type of the primary flare or radio bursts that were used as initial inputs to the HAF model and that are believed to be associated with the listed shock. Where an X-ray flare was observed, its peak magnitude is given. Cases where only a radio burst was observed are labeled as "Type II." Shocks which were not a result of any input other than the background potential source surface magnetic field are labeled as Corotating Interaction Regions (CIRs). The time of the initial forecast availability, (T_{FA}) is also given in Table 1.

3.2. Observations

[25] Real-time (Level 1) data from the Solar Wind Electron, Proton, and Alpha Monitor (SWEPAM) and magnetometer (MAG) instruments aboard the Advanced Composition Explorer (ACE) spacecraft, which is in solar orbit at the L1 point, were used to identify shocks in the solar wind. The SWEPAM and MAG instruments provide measurements of the vector magnetic field at L1, as well as proton radial velocity, density, and temperature at 1-min intervals.

[26] Identifying disturbances based on measurements from a single point in space is quite challenging due to their complex (and largely unknown) morphological nature. Shocks may arrive at L1 head-on, or the disturbance may just graze that point as it heads off in a tangential direction. Additionally, both forward and reverse (those developed downstream) shocks may pass by. Furthermore, some shocks, formed in the vicinity of L1, where the measurements are made, may be transient and short-lived. To study the event-driven shocks that are the focus of this study, it is necessary to isolate them from these other phenomenon using the measurements alone. This is often a somewhat subjective process.

[27] *Kartalev et al.* [2002] describes an automated technique for identifying and classifying shocks from ACE MAG and SWEPAM data using an objective MHD analysis. In this analysis the time series measurements at L1 are separated into upstream and downstream components. If a fundamental change in the MHD characteristics between these two components is observed, then a possible shock is noted and further analysis determines the nature of the shock (e.g., if it is a forward, reverse, or tangential shock, its evolutionarity, and speed relative to the fundamental plasma velocities).

[28] A total of 122 forward shocks were identified by the *Kartalev et al.* [2002] process operating on ACE real-time data for the period. Of these shocks, 24 were removed during a manual inspection due to obvious improper shock identification by the Kartalev et al. algorithm (mostly due to erroneous input data). The remaining 98 shocks are listed by their arrival time at L1 in Table 2.

4. Results

[29] Consistent with the timeline concept discussed in section 2, we evaluate the skill of the HAF forecasts in

Table 1. Solar Wind Shock Arrival Times (T_{FSAT}) From the HAF Model "Fearless Forecasts" for 2001[a]

	T_{FSAT}	*SSI*	T_{FA}	T_0	Event		T_{FSAT}	*SSI*	T_{FA}	T_0	Event
1	2/5/01 1100	0.77	2/4	2/3/01 0000	M2.4	45	8/12/01 1700	0.01	8/11/01	8/10/01 0136	C8.0
2	3/10/01 0600	0.12	3/8	3/8/01 1126	Type II	46	8/17/01 1400	−0.09	8/15/01	8/14/01 1242	C2.3
3	3/13/01 1100	−0.16	3/12	3/10/01 0409	Type II	47	8/23/01 2100	−0.21	8/20/01		CIR
4	3/18/01 0800	−0.08	3/17	3/15/01 2159	C1.9	48	8/25/01 1300	0.08	8/21/01	8/21/01 1024	C2.7
5	3/21/01 0900	−0.29	3/20	3/18/01 0852	C3.1	49	8/27/01 0900	0.23	8/27/01	8/25/01 1632	X5
6	3/24/01 0700	0.07	3/22	3/20/01 0240	M1.1	50	8/30/01 2200	0.47	8/28/01	8/28/01 1603	M1
7	3/25/01 0300	−0.15	3/22	3/20/01 2108	M1.5	51	9/1/01 0900	0.27	8/31/01	8/30/01 0147	C5
8	3/25/01 1800	−0.23	3/22	3/22/01 0822	M1.6	52	9/2/01 0500	0.24	9/1/01	8/31/01 1040	M1.6
9	3/26/01 0600	−0.13	3/25	3/24/01 0139	M1.2	53	9/3/01 0200	0.13	9/1/01	8/31/01 2243	M2.9
10	3/31/01 0300	−0.28	3/27	3/27/01 1632	M2.2	54	9/4/01 2200	−0.26	9/3/01	9/3/01 0158	C9.0
11	3/31/01 0100	0.04	3/30	3/28/01 1240	M4.3	55	9/5/01 1300	−0.34	9/4/01	9/3/01 1832	M2.5
12	4/1/01 2100	−0.02	3/30	3/29/01 1015	X1.7	56	9/8/01 0700	0.27	9/7/01		CIR
13	4/3/01 2200	0.6	4/2	3/31/01 1132	M2.1	57	9/9/01 2300	−0.07	9/7/01		CIR
14	4/12/01 0300	0.22	4/9		CIR	58	9/12/01 2200	0.09	9/11/01	9/9/01 1517	M3.4
15	4/11/01 0400	0.46	4/10	4/10/01 0513	X2.3	59	9/15/01 1500	−0.14	9/13/01	9/11/01 1346	C3.2
16	4/13/01 0800	0.04	4/11	4/9/01 1527	M7.9	60	9/16/01 0600	−0.33	9/13/01	9/12/01 2139	C9.6
17	4/15/01 0300	0.07	4/13	4/11/01 1317	M2.3	61	9/19/01 1600	−0.29	9/15/01	9/15/01 1129	M1.5
18	4/17/01 0600	−0.3	4/13		CIR	62	9/20/01 1500	0.79	9/18/01	9/17/01 0825	M1.5
19	4/21/01 0400	0.06	4/18	4/18/01 0217	C2.2	63	9/23/01 0700	−0.04	9/20/01	9/20/01 0507	C7.5
20	4/25/01 1200	−0.2	4/23	4/22/01 2042	M3.2	64	9/25/01 0200	0.06	9/23/01	9/22/01 0921	Type II
21	4/26/01 1600	0.21	4/27		CIR	65	9/26/01 1000	−0.05	9/25/01	9/24/01 1040	X2.6
22	4/28/01 1600	0.08	4/27	4/26/01 1335	Type II	66	9/30/01 2200	0.26	9/28/01	9/28/01 0830	M3.3
23	5/5/01 1100	−0.16	5/3		CIR	67	10/6/01 1300	0.34	10/2/01		CIR
24	5/6/01 0800	0.06	5/4		CIR	68	10/7/01 0700	−0.3	10/4/01	10/3/01 0647	C6.1
25	5/14/01 0800	−0.02	5/12	5/10/01 1504	Type II	69	10/12/01 0700	−0.16	10/9/01	10/9/01 0737	C7.0
26	5/16/01 0700	0.23	5/15		CIR	70	10/21/01 0600	−0.24	10/19/01	10/19/01 0101	X1.6
27	5/17/01 0100	−0.05	5/15	5/15/01 0300	M1.0	71	10/24/01 0800	0.42	10/22/01	10/22/01 0000	X1.2
28	5/20/01 0500	0.59	5/16	5/16/01 1555	Type II	72	10/27/01 1600	−0.13	10/25/01	10/25/01 1456	X1.3
29	5/22/01 0300	0.37	5/18		CIR	73	11/1/01 1900	−0.07	10/29/01	10/29/01 1113	M3
30	5/23/01 1600	−0.01	5/22	5/20/01 0624	Type II	74	11/6/01 1200	0.17	11/2/01		CIR
31	5/30/01 0600	−0.14	5/26	5/24/01 1940	M.12	75	11/13/01 1900	0.25	11/10/01	11/9/01 1837	M1.9
32	6/1/01 1100	0.12	5/30		CIR	76	11/21/01 1800	−0.08	11/18/01	11/17/01 0450	M2.8
33	6/2/01 1700	0.46	6/1		CIR	77	11/26/01 0800	0.43	11/22/01	11/22/01 2027	M3.8
34	6/7/01 0000	0.12	6/7	6/4/01 0000	Type II	78	12/1/01 0100	0.04	11/28/01	11/28/01 1636	M6.9
35	6/11/01 0700	−0.19	6/8	6/8/01 0000	C6.0	79	12/10/01 1300	0.06	12/7/01		CIR
36	6/14/01 2100	−0.2	6/12	6/11/01 0554	Type II	80	12/11/01 1100	0.21	12/9/01	12/9/01 0443	Type II
37	6/15/01 16:00	−0.14	6/13	6/12/01 0718	Type II	81	12/12/01 0600	−0.1	12/11/01	12/10/01 0940	C8.6
38	6/18/01 0100	−0.03	6/15	6/15/01 1007	M6.3	82	12/13/01 0300	0.33	12/11/01	12/11/01 0808	X2.8
39	6/26/01 2100	0.34	6/22	6/19/01 0335	C1.0	83	12/14/01 0100	0.39	12/10/01		CIR
40	7/3/01 1300	0.31	6/30		CIR	84	12/15/01 1300	−0.11	12/13/01	12/13/01 1429	X6.2
41	7/11/01 0000	−0.29	7/11	7/7/01 0330	C9.0	85	12/22/01 2100	−0.08	12/18/01		CIR
42	7/16/01 2100	−0.26	7/13		CIR	86	12/29/01 1200	−0.32	12/25/01		CIR
43	8/8/01 0500	0.43	8/8		CIR	87	12/28/01 0200	0.11	12/27/01	12/26/01 0502	M7.1
44	8/11/01 0700	0.06	8/10	8/9/01 0936	C3.7						

[a]This list only includes shocks with a Shock Searching Index (SSI) greater than −0.35. Where available, the time of the initiating solar event (T_0) and its type are listed. Those labeled "CIR" correspond to shocks associated with Corotating Interaction Regions that are independent of solar events. T_{FA} corresponds to the first time the shock forecast was available.

terms of the hit window size, ΔT_{SAT}, which will differ from user to user. Table 3 shows the computed value of the skill scores defined in equations (1)–(6) for $\Delta T_{SAT} = 24$, 36, 48, 60 and 72 hours. As can been seen in Table 3, the hit rate H is greater than 60% for all hit window sizes with a trend toward lower values of H as ΔT_{SAT} increases. This is consistent with the fact that H includes all of the correct null forecasts (no shock forecast and no shock observed), which decrease in number as the hit window size increases. The threat score, *TS*, is relatively low, indicating little skill for small ΔT_{SAT}. However, $TS > 0.50$ for $\Delta T_{SAT} > 48$ hours. Also consistent with intuition is the fact that the *POD* and *FAR* metrics monotonically increase (decrease) with ΔT_{SAT} to indicate that the larger the hit window size, the more likely it is to score a correct forecast. The bias, B, is close to 1 for all values of ΔT_{SAT}.

[30] While H, *TS*, *POD*, *FAR*, and B are useful indicators of forecast skill for comparing different forecast models (as is done by *Fry et al.* [2003]) or for tracking long-term trends in forecast performance, they are not particularly useful, by themselves, to the end-user of a forecast because they do not indicate whether the use of a forecast makes economic sense. The $K_{sym} = K_{\theta=1/2}$ skill score, on the other hand, shows the value of the forecast over the alternative optimal naive forecast (the forecast based on the frequency of past events, as described in Appendix A) for cases where the user has a symmetric loss profile (hazard costs equals protection costs). Table 3 shows that $K_{sym} > 0$ for only the 36, 48, and 60-hour hit windows. This indicates that users would be better off not using the HAF forecasts if the requirement is for a hit window less than 36 or greater than 60 hours.

[31] Figure 3 shows the value of the BR skill score, K_θ, as a function of the normalized cost, θ, for the six hit window periods considered. A value of $K_\theta > 0$ indicates that the forecast has skill over the optimal naive forecast. A presentation of K_θ such as shown in this figure is illuminating

Table 2. Forward Shock Arrival Times as Detected From Real Time ACE Data by Kartalev's Objective Algorithm

	T_{OSAT}		T_{OSAT}		T_{OSAT}
1	1/10/01 1523	42	6/21/01 1146	83	10/28/01 0236
2	1/18/01 0852	43	6/22/01 0820	84	10/31/01 1249
3	1/18/01 1129	44	6/23/01 1306	85	12/14/01 1239
4	1/23/01 1002	45	6/24/01 0907	86	12/15/01 1601
5	1/31/01 0723	46	6/25/01 1403	87	12/20/01 1757
6	2/20/01 0053	47	7/23/01 1106	88	12/23/01 2217
7	2/28/01 1324	48	8/2/01 1403	89	12/26/01 0642
8	3/3/01 1036	49	8/3/01 0623	90	12/26/01 0838
9	3/4/01 1156	50	8/9/01 2125	91	12/26/01 0855
10	3/4/01 2236	51	8/12/01 1047	92	12/26/01 0946
11	3/19/01 1026	52	8/14/01 1251	93	12/26/01 1140
12	3/22/01 1239	53	8/16/01 0250	94	12/26/01 1229
13	3/27/01 0108	54	8/16/01 0503	95	12/26/01 1335
14	3/28/01 1552	55	8/16/01 0852	96	12/26/01 1415
15	3/28/01 2257	56	8/16/01 1024	97	12/29/01 0442
16	3/30/01 2152	57	8/17/01 1010	98	12/30/01 1929
17	4/2/01 1945	58	8/17/01 1144		
18	4/3/01 1723	59	8/17/01 1317		
19	4/3/01 1829	60	8/18/01 0522		
20	4/3/01 1903	61	8/22/01 2232		
21	4/4/01 1421	62	8/27/01 1914		
22	4/7/01 1655	63	8/30/01 1332		
23	4/8/01 10:33	64	9/8/01 0520		
24	4/10/01 0002	65	9/8/01 1410		
25	4/10/01 0729	66	9/10/01 0129		
26	4/11/01 1309	67	9/10/01 0243		
27	4/16/01 0336	68	9/24/01 1611		
28	4/17/01 2357	69	9/24/01 1703		
29	4/21/01 1502	70	9/26/01 1400		
30	4/28/01 0428	71	9/26/01 2335		
31	5/1/01 1222	72	9/27/01 0839		
32	5/12/01 1538	73	9/27/01 0937		
33	5/15/01 0302	74	9/30/01 1847		
34	5/18/01 0914	75	10/1/01 2142		
35	5/19/01 0324	76	10/2/01 0225		
36	5/19/01 0450	77	10/2/01 1145		
37	5/30/01 1015	78	10/3/01 1944		
38	6/7/01 0848	79	10/11/01 1613		
39	6/17/01 0424	80	10/15/01 0734		
40	6/18/01 2007	81	10/21/01 1813		
41	6/19/01 1359	82	10/25/01 0806		

because it allows the forecast user to determine the value of a forecast based on the appropriate cost profile. A skill score that is positive for $\theta < 1/2$ is useful for an application where the hazard cost is higher than the protection cost (i.e., unforecasted events are more costly than false alarms). The converse is true for $\theta > 1/2$.

[32] For the symmetric loss case ($\theta = 1/2$), positive skill is indicated for ΔT_{SAT} = 36, 48, and, 60 hours (consistent with Table 3). Greater skill for each of these hit window sizes is also apparent for nonsymmetric loss regimes. For example, the 36-hour case shows the most skill at $\theta = 0.37$ and the 60-hour case peaks at $\theta = 0.57$. The shift of the peak in these curves toward higher θ for longer hit windows is intuitively correct as one would expect that as the hit window size is increased, the forecast is increasingly conservative and fewer missed events are expected at the cost of greater protection. In contrast to Table 3, the 12, 24, and 72-hour hit window forecasts do indeed show skill for various values of θ. In these terms, a figure such as Figure 3 allows decision makers not only to determine the value of a forecast but also to determine the proper forecast to request (in this case in terms of the hit window size).

Table 3. Conventional Skill Scores for the 2001 HAF Forecasted Shock Arrival Times Versus the Kartalev Algorithm Applied to ACE Realtime (Level 1) Data[a]

ΔT_{SAT}	*H*	*TS*	*POD*	*FAR*	*B*	$K_{1/2}$
12	0.79	0.14	0.26	0.74	0.89	−0.42
24	0.67	0.23	0.37	0.63	1.00	−0.26
36	0.67	0.40	0.55	0.45	1.06	0.11
48	0.67	0.46	0.62	0.38	1.14	0.28
60	0.64	0.54	0.66	0.34	1.16	0.16
72	0.65	0.62	0.72	0.28	1.12	−0.22

[a]A $K_{1/2} > 0$ indicates skill.

[33] The significance level of a given value of K_θ can be determined as is shown in BR. The null hypothesis is that $K_\theta \leq 0$. The alternative hypothesis is that $K_\theta > 0$. The appropriate test statistic developed in BR is for cases where $p \leq \theta$ is

$$G_\theta = 2n_{11} \log\left[\frac{\hat{r}}{\theta}\right] + 2n_{01} \log\left[\frac{1-\hat{r}}{1-\theta}\right], \tag{9}$$

where

$$\hat{r} = \frac{n_{11}}{n_{11} + n_{01}}.$$

G_θ has a distribution related to the χ^2 distribution with one degree of freedom. Tests are carried out similar to a standard χ^2_1 test except that the p-value of the ordinary test must be divided by 2 (W. M. Briggs and D. Ruppert, personal communication, 2002); a regular test has that with an observed g, $P(G > g) = \alpha$; this test has that $P(G > g) = \alpha/2$. In practice, the user only has to double the regular χ^2_1 test level to arrive at the correct level and then use an ordinary χ^2_1 test. For example, if the user desires a test level $\alpha = 0.05$,

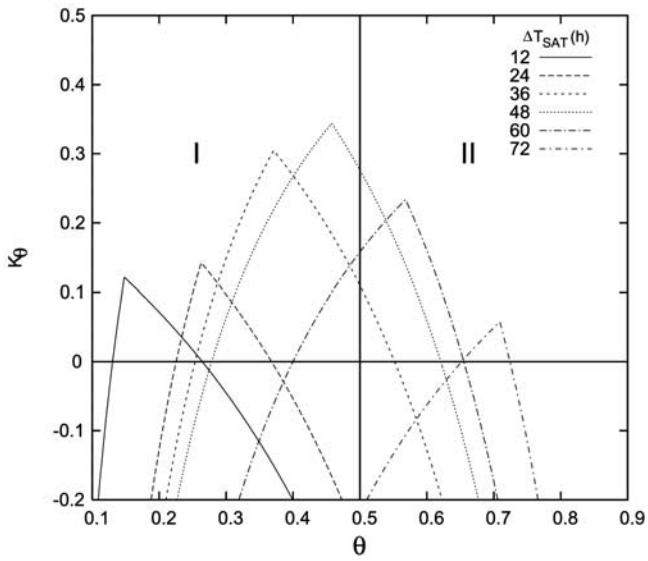

Figure 3. BR forecast skill score, K_θ, as a function of normalized cost, θ, for the six hit-window sizes considered. The quadrant denoted by "I" indicates the region of skill for applications where missed events are more costly than false alarms. Quadrant "II" represents the opposite case.

Table 4. χ^2_1 p-Values Indicating the Statistical Significance of the HAF Shock Arrival Time Forecasts Over a Range of Values of θ for All Six Hit Window Sizes Considered[a]

ΔT_{SAT}	θ = 0.2	θ = 0.3	θ = 0.4	θ = 0.5	θ = 0.6	θ = 0.7
12	0.18					
24		0.21				
36		0.00	0.00	0.37		
48		0.70	0.04	0.0	0.0	
60			1.00	0.21	0.01	
72						0.65

[a]A low p-value indicates high significance in the skill of the HAF forecast over the optimal naive forecast.

he doubles this and uses an ordinary χ^2_1 test with an $\alpha' = 0.10$. This is equivalent to dividing the p-value by 2.)

[34] Table 4 gives the p-value for each of the six hit window sizes as a function of θ. Blank entries in this table indicate no skill over the optimal naive forecast (e.g., $K_\theta \leq 0$). A p-value near zero represents a case where one can be relatively certain that the skill of the forecast relative to the optimal naive forecast is significant and a value near one indicates little certainty of significant skill. A typical significance level would be represented by a p-value around 0.05.

[35] From the data presented in Table 4, it is apparent that we may have confidence in the forecast skill that was indicated in Figure 3 for various combinations of hit window size and normalized cost. In general, the forecasts corresponding to smaller ΔT_{SAT} are significant for smaller values of θ. For the symmetric loss case, only the 48-hour hit window forecast is significant at the 0.05 level. Both the 36 and 48-hour forecasts appear to have significant skill at θ = 0.4 (an application that is slightly more sensitive to missed events than to false alarms). For θ > 0.7, no significant skill exists for any hit window size.

5. Conclusion

[36] The present analysis demonstrates that the current version of the Hakamada-Akasofu-Fry kinematic solar wind model demonstrates some positive skill for forecasting the arrival time of shocks in the solar wind. More importantly, however, we have demonstrated a new paradigm for evaluating forecast skill for the space weather community that incorporates aspects of the particular application of the forecast. By including the relative costs to the user of missed forecasts, one is able to couch the relative benefits of using the forecast, such as through the use of the Briggs-Ruppert skill score presented here.

[37] The graphical methods and statistical testing of skill introduced here will allow a wide range of decision makers to easily use and interpret the solar shock forecasts.

[38] Future improvements to heliospheric forecast models such as HAF, as well as improved observational data, such as is expected from SMEI, offer the potential to greatly improve the inherent skill in space weather forecasts and to go beyond simple shock arrival time predictions to forecasts of impacts on space-based and other sensitive systems. Analysis of the value of these forecasts in terms of the economic benefits associated with the specific application of these forecasts, such as presented here, will be fundamental to their evaluation and eventual use in operational contexts.

Appendix A: Briggs Ruppert Skill Score

A1. Definitions

[39] We are concerned with events Y which are dichotomous. Predictions (which may be probabilistic) $X \in [0, 1]$ are made for Y. In the two decision problem a decision maker acts on X in one of two ways: he takes action d_1 if he believes Y will occur and d_0 if he believes it will not. Predictions can be either dichotomous or probabilistic, but here we only consider decisions that are dichotomous. This implies a transformation of a probabilistic prediction into an eventual dichotomous one. In this work only the transformed or purely dichotomous prediction is used. The method of transformation is shown below.

[40] We follow the notation developed by *Schervish* [1989]. Let $Y_i \in \{0, 1\}$ designate the ith observation of a dichotomous event; that is, $Y_i = 1$ if the event occurs and equals zero if it does not. Let the loss k associated with making a correct decision equal 0. The loss k for making an error can always be quantified such that when $Y = 0$ and the decision was d_1 it is θ, which implies that when $Y = 1$ and the decision was d_0 the loss was $1 - \theta$.

[41] The decision maker minimizes his loss and takes d_1 whenever (the possibly probabilistic) expert forecast $X_i \geq \theta$ or takes d_0 if $X_i < \theta$. The loss for any given forecast can be now written as

$$k_i = \theta I(X_i \geq \theta, Y_i = 0) + (1 - \theta) I(X_i < \theta, Y_i = 1). \quad \text{(A1)}$$

[42] Skill will be framed in terms of expected loss (risk). In order for a collection of predictions to have skill, we desire that its expected loss should be less than the expected loss incurred by using the optimal naive predictions (typical definitions of skill, for example by *Wilks* [1995], refer to skill as relative accuracy of an expert to a naive forecast, a distinction we will keep when developing the skill score below). The naive information we have about Y is p, the unconditional probability of occurrence, so that skill, if it exists, is known as "climate" or "simple skill" to reflect the idea that the expert forecast can beat the climatological or simple forecast. The expected loss for a decision maker for a collection of expert predictions is

$$\begin{aligned} E(k) &= E(\theta I(X \geq \theta, Y = 0) + (1 - \theta) I(X < \theta, Y = 1)) \\ &= \theta P(X \geq \theta, Y = 0) + (1 - \theta) P(X < \theta, Y = 1) \\ &= \theta P(\tilde{X} = 1, Y = 0) + (1 - \theta) P(X = 0, Y = 1). \end{aligned} \quad \text{(A2)}$$

The last step uses the fact that $P(X \geq \theta, Y = 0) = P(\tilde{X} = 1, Y = 0)$, where $\tilde{X}$ reflects the (possible) probabilistic prediction transformed to a dichotomous one by the decision maker. Unless otherwise indicated, the tilde shall be dropped and it shall be assumed that the (transformed if necessary) dichotomous predictions are used (this transformation differs from that of *Mason* [1979] whose goal was to maximize the expected value of various scores that are based on the forecast).

[43] Thus, let $X_i \in \{0, 1\}$ designate the ith prediction. Let $P(X = 1) = q$, $P(Y = 1) = p$. Further let $P(Y = 1 | X = 1) = r$ and $P(Y = 1 | X = 0) = s$. We also assume that each observation Y_i

is independent of each other Y_j for $i \neq j$ and that all of these probabilities are unvarying for all i. We also assume that $cov(Y_i, X_j) = 0$ for $i \neq j$; that is, the forecast observation process is not dynamic and that future observations do not depend on past forecasts (for extensions to Markov Y see W. M. Briggs and D. Rupert, personal communication, 2002).

[44] The expected loss for the optimal naive forecast depends both on p and on the value of θ. If $p \leq \theta$ the optimal naive forecast is $X^N = 0$, where the superscript X^N denotes the naive forecast. This gives an expected loss of $p(1 - \theta)$. If $p > \theta$ the optimal naive forecast is to always answer $X^N = 1$. The expected loss is $(1 - p)\theta$. It is convenient in what follows, but not necessary, to transform both the observations and the loss so that the optimal naive prediction is always $X^N = 0$. The transformation is $Y' = 1 - Y, X' = 1 - X$, $\theta' = 1 - \theta$, which gives $p' = 1 - p$. The transformation, if any, is only done to ensure $p \leq \theta$. This transformation simplifies the presentation but does not change any results.

A2. Skill Test

[45] The null hypothesis for the skill test can now be formed. It is

$$H \quad E(k^E) \geq E(k^N) \tag{A3}$$

where k^E corresponds to the loss of the expert prediction and k^N is the loss of the optimal naive prediction, and expectation is taken over both forecasts and observations.

[46] Note that $p = P(Y = 1, X = 1) + P(Y = 1, X = 0)$. Substituting for the expected loss gives the null hypothesis

$$\theta P(Y = 0,\ X = 1) + (1 - \theta)P(Y = 1,\ X = 0) \geq p(1 - \theta)$$

$$\begin{aligned} \theta P(Y = 0, X = 1) &\geq P(Y = 1, X = 1)(1 - \theta) \\ \theta &\geq \frac{P(Y = 1, X = 1)}{q} \\ r &\leq \theta \end{aligned} \tag{A4}$$

[47] The alternative is $r > \theta$ (note that in those cases where $p > \theta$ and the user has opted not to recast the forecasts and observations, the null translates to $s \geq \theta$).

[48] Assume, for the moment, that the loss is symmetric; that is, $\theta = 1/2$. An alternative interpretation of skill requires the probability that $Y = X^E$ exceeds the optimal naive probability of $Y = X^N$ (or $Y = 0$), or, for the null,

$$\begin{aligned} H \quad P(Y = X) &\leq P(Y = 0) \\ rq + (1 - s)(1 - q) &\leq (1 - r)q + (1 - s)(1 - q) \\ r &\leq 1/2. \end{aligned}$$

This is identical to the original null with $\theta = 1/2$.

[49] The likelihood of the model, written in terms of r, s, q, is

$$\begin{aligned} L(r, s, q|Y, X) = \prod_{i=1}^{n} & q^{X_i}(1 - q)^{1 - X_i} r^{X_i Y_i}(1 - r)^{X_i(1 - Y_i)} \\ & \times s^{(1 - X_i)Y_i}(1 - s)^{(1 - X_i)(1 - Y_i)}. \end{aligned} \tag{A5}$$

The estimates for these parameters are found in the cell counts of a 2×2 observation and prediction contingency table shown in Figure 2.

[50] The unrestricted maximum likelihood estimates (MLEs) are easily found as each parameter separates in the likelihood:

$$\begin{aligned} \hat{q} &= \frac{n_{11} + n_{01}}{n} \\ \hat{r} &= \frac{n_{11}}{n_{11} + n_{01}} \\ \hat{s} &= \frac{n_{10}}{n_{10} + n_{00}}, \end{aligned}$$

where $n = \Sigma_i \Sigma_j n_{ij}$. Under the null the MLE for q remains unchanged as might be expected as it only involves the unconditional mean of the forecast X. The null is that $r \leq \theta$, maximized at $r = \theta$, and the estimate for s remains unchanged. These facts makes calculation of the likelihood ratio statistic (LRS) particularly simple as the terms involving q and s drop out, leaving only the terms involving r.

[51] The LRS, G, is

$$\begin{aligned} G &= -2 \log \left[\left(\frac{\theta}{\hat{r}} \right)^{n_{11}} \left(\frac{1 - \theta}{1 - \hat{r}} \right)^{n_{01}} \right] \\ &= 2n_{11} \log \left[\frac{\hat{r}}{\theta} \right] + 2n_{01} \log \left[\frac{1 - \hat{r}}{(1 - \theta)} \right]. \end{aligned}$$

As a practical matter, when making calculations with real data the often used condition $0 \log(0) = 0$ is invoked.

[52] G has an asymptotic distribution which is related to the χ^2 distribution with one degree of freedom. Since the test is one-sided the actual distribution is $1/2\chi^2_1 + 1/2o_p(1)$ [*Cox and Hinkley*, 1974]. Tests are carried out similar to a standard χ^2_1 test, except that where a normal χ^2_1 statistic W has that $P(W > w) = \alpha$, here because there is a probability mass of 1/2 at 0, the χ^2_1 statistic G has that $P(G > w) = \alpha/2$. In practice, the user only has to double his chosen test level and use an ordinary χ^2_1 distribution.

A3. Skill Score

[53] Those interested in forecast evaluation typically want not only to know whether a collection of forecasts has been skillful, but they would also like to attach a number or score that measures this skill. This is useful, for example, in tracking skill for a system of forecasts over time or for comparing forecasts made for similar events. Skill scores are in widespread use in the meteorological community [*Wilks*, 1995; *Kryzysztofowicz*, 1992]. Normally, skill scores K take a form such as the following:

$$K(y, x) = \frac{S(y, x^N) - S(y, x^E)}{S(y, x^N)}, \tag{A6}$$

where $S(y, x^N)$ is an error score for a collection of naive forecasts, and $S(y, x^E)$ is the same error score for a collection of expert forecasts. The divisor is there to "normalize" the error scores so that rough comparisons can be made between skill scores received across different situations. Scores of the type of equation (A6) are not proper [*Winkler*, 1996]. A proper score is one in which $E_p(K(y, p)) \geq E_p(K(y, x))$, and reflects the idea that the forecaster can only maximize his score by forecasting his true feeling

[*Hendrickson and Buehler*, 1971]. Winkler shows a proper score related to equation (A6) is $K(y, x) = S(y, x^N) - S(y, x^E)$; however, this loses the desirable normalizing quality of equation (A6). Scores in the form of equation (A6) are in widespread use, and the departure from properness does not do much harm here (in the sense that a forecaster will find it difficult if not impossible to manipulate the eventual skill score to his advantage, and *Murphy* [1973] has also shown that skill scores of this form are approximately proper for large samples).

[54] The difficulty with skill scores has traditionally been that the sampling distribution of the skill score was unknown making hypothesis testing impossible. However, in our case testing the significance a skill score is the same as the normal skill test if the following skill score is taken:

$$K(y, x^E) = \frac{E(k(Y, X^N)) - E(k(Y, X^E))}{E(k(Y, X^N))}, \tag{A7}$$

where the expected forecast loss is taken as the error score. A collection of perfect expert forecasts will have a loss of 0, so, for us, a perfect skill score will be $K \equiv 1$. A collection in which no skill exists will have either an expected loss the same as the naive forecasts or even greater so that the skill score will be 0 or less. The null hypothesis is

$$H \quad K \leq 0. \tag{A8}$$

It can be easily seen that this translates exactly to the hypothesis and test used before.

[55] An estimate for the skill score is

$$\begin{aligned}\widehat{K}_\theta &= \frac{\widehat{p}(1-\theta) - \theta(1-\widehat{r})\widehat{q} - (1-\theta)\widehat{s}(1-\widehat{q})}{\widehat{p}(1-\theta)} \\ &= \frac{(\widehat{r}-\theta)\widehat{q}}{\widehat{p}(1-\theta)} \\ &= \frac{n_{11}(1-\theta) - n_{01}\theta}{(n_{11}+n_{10})(1-\theta)}.\end{aligned} \tag{A9}$$

[56] **Acknowledgments.** The authors would like to acknowledge Michelle Rooney for her assistance in data analysis. We would also like to thank Murray Dryer, Zdenka Smith, Tom Detman, and Ghee Fry for their useful input into this work. This work was supported in part by the Air Force Office of Scientific Research and the Air Force Research Laboratory's Space Scholars Program.

[57] Shadia Rifai Habbal thanks Alan W. P. Thomson and Murray Dryer for their assistance in evaluating this paper.

References

Cox, D. R., and D. V. Hinkley, *Theoretical Statistics*, John Wiley, New York, 1974.

Dryer, M., and D. F. Smart, Dynamical models of coronal transients and interplanetary disturbances, *Adv. Space Res.*, *4*, 291–301, 1984.

Fry, C. D., W. Sun, C. S. Deehr, M. Dryer, Z. Smith, S.-I. Akasofu, M. Tokumaru, and M. Kojima, Improvements to the HAF solar wind model for space weather predictions, *J. Geophys. Res.*, *106*, 2985–2991, 2001.

Fry, C. D., M. Dryer, Z. Smith, W. Sun, C. S. Deehr, and S.-I. Akasofu, Forecasting solar wind structures and shock arrival times using an ensemble of models, *J. Geophys. Res.*, *108*(A2), 1070, doi:10.1029/2002JA009474, 2003.

Gombosi, T. I., D. L. DeZeeuw, C. P. T. Groth, K. G. Powell, C. R. Clauer, and P. Song, From Sun to Earth: Multiscale MHD simulations of space weather, in *Space Weather*, *Geophys. Monogr. Ser.*, vol. 125, edited by P. Song, H. J. Singer, and G. L. Siscoe, p. 169, AGU, Washington, D. C., 2001.

Hendrickson, A. D., and R. J. Buehler, Proper scores for probability forecasters, in *Ann. Math. Stat.*, *42*, 916–920, 1971.

Kartalev, M. D., K. G. Grigorov, Z. Smith, M. Dryer, C. D. Fry, W. Sun, and C. S. Deehr, Comparitive study of predicted and experimentally detected interplanetary shocks, in Proc. SOLSPA: The Second Solar Cycle and Space Weather Euroconference, SP-477 Eur. Space Agency, Paris, 2002.

Kryzysztofowicz, R., Bayesian correlation score: A utilitarian measure of forecast skill, *Mon. Weather Rev.*, *120*, 208–219, 1992.

Luhmann, J. G., CMEs and space weather, in *Coronal Mass Ejections*, *Geophys. Monogr. Ser.*, vol. 99, edited by N. Crooker, J. A. Joselyn, and J. Feynman, pp. 291–299, AGU, Washington, D. C., 1997.

Mason, I., On reducing probability forecasts to Yes/No forecasts, *Mon. Weather Rev.*, *107*, 207–211, 1979.

Murphy, A. H., Hedging and skill scores for probability forecasts, *J. Appl. Meteorol.*, *12*, 215–223, 1973.

Riley, P., J. Linker, Z. Mikić, and R. Lionello, MHD modeling of the solar corona and inner heliosphere: Comparison with observations, in *Space Weather*, *Geophys. Monogr. Ser.*, vol. 125, edited by P. Song, H. J. Singer, and G. L. Siscoe, pp. 159, AGU, Washington, D. C., 2001.

Schervish, M., A general method for comparing probability assessors, *Ann. Stat.*, *17*, 1856–1879, 1989.

Smart, D. F., and M. A. Shea, A simplified technique for estimating the arrival time of solar flare-initiated shocks, in *Proceedings of STIP Workshop on Solar/Interplanetary Intervals*, edited by M. A. Shea, D. F. Smart, and S. McKenna-Lawlor, pp. 139–156, Book Crafters, Chelsea, Mich., 1984.

Smart, D. F., and M. A. Shea, A simplified model for timing the arrival of solar-flare-initiated shocks, *J. Geophys. Res.*, *90*, 183–190, 1985.

Smith, Z., and M. Dryer, MHD study of temporal and spatial evolution of simulated interplanetary shocks in the ecliptic plane within 1 AU, *Sol. Phys.*, *129*, 387–405, 1990.

Smith, Z. K., M. Dryer, E. Ort, and W. Murtagh, Performance of interplanetary shock prediction models, *J. Atmos. Sol. Terr. Phys.*, *62*, 1264–1274, 2000.

Sun, W., M. Dryer, C. D. Fry, C. S. Deehr, Z. Smith, S.-I. Akasofu, M. D. Kartalev, and K. G. Grigorov, Real-time forecasting of ICME shock arrivals at L1 during the "April Fool's Day" epoch: 28 March–21 April 2001, *Ann. Geophys.*, *20*, 937–945, 2002.

Thomson, A. W. P., Evaluating space weather forecasts of geomagnetic activity from a user perspective, *Geophys. Res. Lett.*, *27*, 4049–4052, 2000.

Webb, D. F., J. C. Johnston, and R. R. Radick, The Solar Mass Ejection Imager (SMEI): A new tool for space weather, *Eos Trans. AGU*, *83*, 38–39, 2002.

Wilks, D. S., *Statistical Methods in the Atmospheric Sciences*, 467 pp., Academic, San Diego, Calif., 1995.

Wilks, D. S., A skill score based on economic value for probability forecasts, *Meteorol. Appl.*, *8*, 208–219, 2001.

Winkler, R. L., Scoring rules and the evaluation of probabilities (with comments), *Test*, *5*, 1–60, 1996.

W. M. Briggs, Weill Medical College, Cornell University, 525 E. 68th Street, Box 46, New York, NY 10021, USA. (wib2004@med.cornell.edu)

J. B. Mozer, Space Weather Center of Excellence, Space Vehicles Directorate, Air Force Research Laboratory, P.O. Box 62, Sunspot, NM 88349, USA. (jmozer@nso.edu)