ADVANCING EARTH AND SPACE SCIENCE

Space Weather

RESEARCH ARTICLE

10.1029/2020SW002604

Key Points:

- Skill scores of forecast models use only numbers of events
- Model validations ignore sizes or impacts of events
- We present an intensity-based skill score for models

Correspondence to:

S. W. Kahler,
stephen.kahler@us.af.mil

Citation:

Kahler, S. W., & Darsey, H. (2021). Exploring contingency skill scores based on event sizes. *Space Weather*, *19*, e2020SW002604. https://doi.org/10.1029/2020SW002604

Received 5 AUG 2020
Accepted 2 APR 2021

Exploring Contingency Skill Scores Based on Event Sizes

S. W. Kahler[1] **and H. Darsey**[2]

[1]Air Force Research Laboratory, Kirtland AFB, Albuquerque, NM, USA, [2]University of New Mexico, Albuquerque, NM, USA

Abstract Space weather forecasts are generally made for events with an arbitrary size threshold imposed on an event statistical size distribution which is likely described by a power law. This is the case for solar energetic ($E > 10$ MeV) particle (SEP) events, which have a differential power law exponent of $\gamma = 1.2$. Event forecasts are usually evaluated by skill scores using a contingency table that matches the forecasted events against observed events independently of the event sizes. Each observed event is either a forecasted hit or a miss, and each forecasted event is either an observed hit or a false alarm. However, for SEP events and most other space weather parameters the event size is a critical factor for the user. It is more important that large events be well forecasted than threshold events. In addition, false alarms may be useful when they match observed events just below the forecast threshold. We explore a forecast evaluation scheme to incorporate the event size within the usual format of a binary contingency table to evaluate model performance. The scheme is applied to three different input options of a recently published evaluation of the Proton Prediction System (PPS) for SEP events to show differences between numbers-based and intensity-based skill scores of the PPS. We demonstrate how identical skill scores can result from models with extremely different performances of event intensity forecasts. The scheme requires model validation and would benefit from testing with other space weather applications.

1. Introduction

Space weather is now recognized to have important deleterious effects on human health and technology (Chiarini, 2013; Schrijver et al., 2014, 2015; Tobiska et al., 2015; Zheng et al., 2019). Increasing efforts are devoted to various models for forecasting the occurrence and magnitudes of solar transient events such as flares (Barnes et al., 2016; Bobra & Couvidat, 2015), coronal mass ejections (CMEs) (Bobra & Ilonidis, 2016; Falconer et al., 2014; Verbeke et al., 2019), and solar energetic particle (SEP) events (Marsh et al., 2015; Núñez et al., 2019; Papaioannou et al., 2018; Richardson et al., 2018; Zhong et al., 2019), that are the principal drivers of most space weather (Schrijver & Siscoe, 2010). The forecast models are usually based on either statistical compilations of archival solar events, assumed physical models, or both. The verification of any forecast model is its application to historical event databases. The goal then becomes to show via a quantitative metric the merits of the forecast model.

Forecast verification is also important for terrestrial weather forecast models. As reviewed by Casati et al. (2008), the field faces challenges to define verification procedures that address user requirements dealing with spatial structures and the presence of features in forecast fields. Ensemble forecasts have motivated the development of new verification methods for probability distribution functions, and the incorporation of operational feedback into forecast models and correct interpretation of verification statistics are active areas of research. Work on forecast verification methods is the topic of a number of ongoing workshops, conferences, and demonstration projects.

Research on extreme events, driven by their socio-economic impacts, is a particular focus of the community. Understanding, modeling, and predicting weather and climate extremes is the goal of the Extremes Grand Challenge of the World Climate Research Program (Sillman et al., 2017). Scientific challenges are to understand the large-scale drivers and regional feedback processes to improve prediction methods and to assess the model performances. Two classes of extreme events are divided by time scales of short duration (<3 days, e.g., tornados, lightning, storm surges, cyclones, and anticyclones) and long duration (>3 days, e.g., droughts, heatwaves, floods, and increased wildfire seasons). The urgent need is for better observations and model evaluation tools specifically suited to the analysis of extremes. Evaluation of model ensembles and their properties or specific features, such as mean or variance, is a basic concern. The public attention paid to catastrophic extreme event forecasts and the restriction of evaluation to subsets of available forecasts can mean that skillful forecasts are unfairly discredited (the forecaster's dilemma, Lerch et al., 2017). How one might apply scoring rules to probabilistic forecasts when the particular emphasis is placed on extreme events while retaining propriety is a fundamental question. From the perspective of proper scoring rules, restricting the outcome space corresponds to the multiplication of a scoring rule by an indicator weight function, which renders any proper score improper. The quest for terrestrial weather (and other) communities is to find suitably weighted scoring rules for probabilistic models that allow for emphasis on extreme events.

Table 1
Basic Contingency Table for Deterministic Forecasts of a Sequence of n Binary Events Showing Numbers of Observed and Forecasted Events

Observed				
		Yes	No	
Forecast	Yes	a	b	a + b
	No	c	d	c + d
		a + c	b + d	n

Note. a = number of events both forecasted and observed. b = number of events forecasted but not observed. c = number of events not forecasted but observed. d = number of events not forecasted and not observed. n = total number of events.

While extreme climatological events have continued to increase over the past several decades, extreme space weather events have trended in the opposite direction, toward fewer and weaker extreme events. Solar cycle (SC) 24, which ended about December 2019, was characterized by lower sunspot numbers (SSNs) than in SC 23 (1996–2008), which was lower than those of SC 22 (Ahluwalia, 2019). A lower SC 24 heliospheric pressure resulted in slower CMEs of decreased magnetic energy and geoeffectiveness (Gopalswamy et al., 2020). SEP ground level events (GLEs) of SC 24 decreased to 2 from 16 in SC 23, and Dst ≤ -100 nT geomagnetic storms to 22 from 86. A forecast for the maximum smoothed monthly SSN of SC 25 (Svalgaard, 2020) is 128 ± 10, only slightly higher than the SC 24 maximum of 116. A phenomenological model prediction of McIntosh et al. (2020), however, yields a preliminary value of 233, with a 68% confidence range of 204–254, far above the Solar Cycle 25 Prediction Panel consensus maximum of 95–130.

In this work, we briefly review the standard tools and then suggest a new method to incorporate an event intensity of user importance into a forecast. The intensity-based and standard number-based forecasts are contrasted within the context of existing verification methods.

Space weather forecasts can be done either as continuous predictands, that is, probabilistic forecasts, or as categorical forecasts of discrete predictands (Stephenson, 2000; Wilks, 2006). The simplest example of the latter are forecasts of the occurrence or absence of a particular event for a specified time. The forecast can then be compared with the subsequent observation to determine whether the event forecast was correct or not. Contingency tables in a 2×2 format are a standard method of evaluating space weather event forecasts (Wilks, 2006), such as the occurrence of solar energetic particle (SEP) events (Zhong et al., 2019) or geomagnetic storms (Jackson et al., 2019). Table 1 shows the basic format where a is the number of events both forecasted and observed (hits), b is the number forecasted but not observed (false forecasts), c is the number observed but not forecasted (misses), d is the number of correctly forecasted null events, and the total n = a + b + c + d.

Several basic measures of the model are probability of detection

$$POD = \frac{a}{a + c}, \tag{1}$$

false-alarm ratio

$$FAR = \frac{b}{a + b}, \tag{2}$$

and bias

$$B = \frac{a + b}{a + c}, \tag{3}$$

which is 1 when the forecasted events equal the observed events. These and other measures are defined and discussed in Chapter 7 of Wilks (2006).

Skill scores are scalar measures to determine the accuracy of a given model relative to some reference model (Barnes et al., 2016; Wilks, 2006). A commonly used measure of relative model accuracy based on contingency tables is the Heidke skill score (HSS), which can be written as:

$$HSS = \frac{2(ad - bc)}{(a + c)(c + d) + (a + b)(b + d)}, \tag{4}$$

when the model is compared to random forecasts with the same probabilities of forecasted events (a + b)/n and of observed events (a + c)/n as in the model (Wilks, 2006).

A second kind of skill score, the TSS, or true skill statistic, compares the model with a reference unbiased ($B = 1$) model and is given by

$$TSS = \frac{ad - bc}{(a + c)(b + d)}. \tag{5}$$

An advantage of the TSS over that of the HSS is that TSS is unchanged between model validations with different data sets a and c, but the same PODs, whereas the HSS, in general, is different (Bloomfield et al., 2012).

The HSS and TSS serve as good quantitative scalar measures of the values of forecast models, particularly when competing models must be judged in selection for implementation. We have seen that these criteria measure accuracy relative to naive forecasts. In both cases, the measures range from one for perfect forecasts to 0 for random forecasts to negative values for worse than random forecasts. They are widely used as basic tools in space weather forecast evaluations (e.g., Barnes et al., 2016; Inceoglu et al., 2018; Zhong et al., 2019).

The TSS and HSS scores reflect the relative performance of space weather models well when all observed events have the same intrinsic value to the user. In that case, there are no large or small events in terms of their space weather application, and all events are considered equally. One example is the forecasting of whether observed CMEs will arrive at Earth (Verbeke et al., 2019). Further, false alarms should be equal in their consequences to missed events. This is evident by exchanging b and c in the HSS above and noting that TSS is evaluated for the condition of b and c equivalency. While Equations 2 and 3 provide a convenient method of model evaluations, users in space weather situations almost always differ in their tolerance of false alarms versus missed events, so the question is how to account for deviations from those optimal conditions. Mozer and Briggs (2003) addressed the problem of equivalency of b and c by developing an economic skill score for categorical forecasts proposed by Briggs and Ruppert (BR, private comm.) and generalized from the work of Thomson (2000). The basic idea is that the cost of a false positive forecast, c_{01}, may not equal the cost of a false negative forecast, c_{10}, which is captured by

$$\theta = \frac{c_{01}}{(c_{01} + c_{10})}. \tag{6}$$

Mozer and Briggs (2003) plotted their BR skill scores for a solar wind shock forecast model for $0.1 < \theta < 0.9$.

In the terrestrial weather community, Murphy (1985) considered a general model of Wj degrees of adverse weather for which the user can take Pi levels of protection, where j and i range from 1 (completely adverse weather and full protection) to N (no adverse weather and no protection). For combinations of i > j, a loss is incurred, and for j > i, an additional cost results. Based on several assumptions about the actions taken, he derived expressions for the expected expenses incurred for imperfect, climatological, and perfect forecasts. For the reduced case of N = 2, of interest here, Wilks (2001) proposed a forecast evaluation to address the cost/loss ratio problem, in which a decision-maker can either pay a cost C to protect against effects of adverse weather or suffer a loss L when no protection is taken and adverse weather occurs. He defined a Value Score VS in terms of C/L and a, b, and c of Table 1, interpreted as the expected economic value of forecasts as a fraction of the value of perfect forecasts relative to climatological forecasts.

In the space weather community, Park et al. (2017) addressed the cost/loss problem for forecasts of three classes of solar flares and defined a Value Score with a formula identical to that of Wilks (2001). Owens and Riley (2017) compared deterministic versions of an MHD model of Vsw with large ensembles of Vsw produced by the same model. They evaluated the two methods using the cost/loss model of Murphy (1985) where the user takes an action with cost C or risks a loss L for various threshold values of Vsw. The use of model ensembles and cost/loss analysis has been lauded by Henley and Pope (2017) as welcome adaptations of terrestrial weather forecasting techniques to space weather forecasting. These models are concerned with the consequences of particular responses to forecasts, but continue to distinguish only between individual cases of forecasts and no forecasts.

A more important skill score limitation is that of event size equivalency, in which all events are treated as being of equal importance, or size, to the user of evaluated models. Two extreme situations challenge that concept, the first being one of two events with huge differences in size or magnitude. When the forecasted parameter can range over orders of magnitude, a small event barely above the threshold is equivalent in the skill scores to another larger by orders of magnitude, in that both may fall in the same category of Table 1. At least for those two hypothetical events, the model that fails to forecast the huge event while correctly forecasting a threshold event merits the same skill score contribution as the one that forecasts the former and misses the latter because both events are given the same weight.

The second extreme situation is one in which models must distinguish between two nearly identical events near the forecast event threshold. For a forecast event threshold of say 100 units, the model forecasting a 105-unit event but not a 98-unit event is deemed superior to a model missing the first but forecasting the second. Thus, forecast models based on thresholds can fail to provide event size resolution when needed and impose it when it is essentially useless. These threshold-based issues are often not present in some terrestrial weather forecasts, where all storms or tornados might be considered equivalent and count the same. It then does not matter which tornado was missed or falsely forecasted—they are all the same for an evaluation score.

2. Size Distributions and Weighting of Events

It is important to consider the parent population of events to be forecasted by a given model. In space weather, many event differential size distributions, such as those of SEP peak intensities Ip (Belov et al., 2007; Cliver & D'Huys, 2018) and solar flares (Aschwanden, 2019; Cliver & D'Huys, 2018; Ryan et al., 2016) are described by power laws, but the power-law fits, or even the appropriate fitting functions, can be misleading when based only on graphical methods employed with small sample sizes (Cliver & D'Huys, 2018; Verbeeck et al., 2019). The functional uncertainty can become acute when events in the distribution tail dominate a radiation model (Jiggens et al., 2018). Other space weather distributions, such as geomagnetic Dst, are described by exponentials (Echer et al., 2011) or by Weibull functions (Gopalswamy, 2017).

In each of these distributions, there are many small events, each with little user impact, and a few big events, each with major impacts. Depending on the application, the range of forecasted variable intensities can be several orders of magnitude, almost universal for space weather variables. The forecast threshold should be set to include the many small events, which cumulatively may have an impact comparable to or exceeding that of the few large events. On the other hand, the included small events may play only a cumulative minor role compared to the impacts of the few large events. Providing a resolution to this uncertainty is the motivation here for exploring a new system of event-weighted variables. We have seen above that the standard skill scores based only on numbers of event detections fail to capture that size importance.

To illustrate our event-weighted evaluation, we need a forecast parameter with a significantly declining size distribution, although not necessarily a power law. Next is a set of event data and several competing forecast models which can be evaluated with standard skill scores in addition to our weighted method. We note that many published forecast model evaluations not only do not list their input observations, which may be too extensive for publication but give only their skill scores without the input contingency tables (e.g., Zhong et al., 2019).

Table 2
$E > 50$ MeV SEP Events and PPS Forecast Contingency Table Elements for Three Input Groups of KWL17

			> M5		
Peak	UT	Ip	flare	> M5 flare	8800 MHz
Date	Time	(p/cm^2 s sr)	Only	& 8800 MHz	> 500 sfu
2/7/1986	17:50	14.8	c	a	a
2/14/1986	17:45	9.35	e	b	b
5/4/1986	12:55	1.53	e	e	e
1/3/1988	8:35	6.07	e	e	e
3/25/1988	22:35	3.63	e	e	e
11/8/1988	21:05	1.73	e	e	e
12/15/1988	3:00	1.71	e	e	e
12/16/1988	22:50	2.89	d	e	e
3/12/1989	19:25	183	c	a	a
3/13/1989	7:45	10	c	c	c
3/18/1989	8:15	10.3	a	a	a
3/23/1989	21:00	2.36	d	b	b
6/18/1989	17:10	3.52	e	e	e
7/25/1989	10:40	17.5	c	c	c
8/13/1989	4:40	557	a	a	a
8/23/1989	21:25	1.27	e	e	e
9/30/1989	6:20	5,650	a	a	a
10/20/1989	10:20	5,150	c	a	a
10/22/1989	23:30	700	a	a	a
10/24/1989	23:00	500	a	a	a
10/29/1989	10:00	7	e	e	e
11/15/1989	8:30	12.5	a	a	a
12/1/1989	10:45	42.4	a	a	a
3/19/1990	16:55	5.68	b	b	e
4/28/1990	17:20	3.81	e	e	e
5/22/1990	5:00	61.5	a	a	a
5/24/1990	22:00	43.4	a	a	a
7/26/1990	7:30	1.23	e	e	e
8/1/1990	17:05	1.59	e	e	e
3/24/1991	3:50	2,510	c	c	c
5/13/1991	4:40	24.6	c	a	c
6/8/1991	17:30	23.2	c	c	c
6/11/1991	13:50	269	a	a	a
6/15/1991	13:00	153	a	a	a
6/30/1991	17:15	1.84	e	e	e
7/1/1991	20:55	2.5	d	e	e
7/8/1991	6:45	4.79	e	e	e
10/30/1991	11:10	9.93	d	b	b
3/8/1992	9:10	2.18	e	e	e

3. Data Analysis: The Proton Prediction System

The proton prediction system (PPS) (Kahler et al., 2007, 2017) provides deterministic forecasts of $E > 10$ MeV proton events observed with the GOES energetic proton sensors for peak intensities Ip of ≥ 10 proton flux units (pfu, 1 p cm^{-2} s^{-1} sr^{-1}), using solar flare longitudes and either X-ray flare or radio burst observations as forecast model inputs. The PPS model first calculates SEP time-intensity profiles in 18 integral energy ranges to estimate a forecasted peak intensity in each range and then produces a yes/no forecast based on whether that intensity exceeds a specified threshold of 10 pfu. It does not forecast the calculated peak intensity. After Kahler et al. (2007) validated PPS for $E > 10$ MeV protons with GOES ≥ M5 flare peak flux and flare fluence (flux × rise time) inputs, Kahler et al. (2017) (KWL17) extended the PPS validation to $E > 50$ MeV protons, an energy range considered more important for space weather applications. As forecast model test inputs, KWL17 used associated ≥ M5 X-ray flare fluences and three different groups of peak fluxes of 8800-MHz radio bursts: (1) 8800-MHz bursts accompanied by ≥ M5 flares; (2) all 8800-MHz bursts > 500 solar flux units (sfu) and known longitude sources; and (3) all 8800-MHz bursts > 5,000 sfu with known longitude sources. Contingency tables for all four input groups to the PPS are given in Table 2 of KWL17. Their PPS results will be the basis of our examples of weighting events for skill score evaluations. We exclude the fourth group of all 8800-MHz radio bursts with peak fluxes > 5,000 sfu and known flare longitudes, which was limited to 81 flares and produced the worst results of the four groups. KWL17 focused on a target population of 67 10-pfu threshold $E > 50$ MeV SEP events over the period 1986 to 2016, and, following the example of Kahler et al. (2007), included a separate population of 71 SEP events with Ip between 1 and 10 pfu.

After compiling contingency tables for the Ip ≥ 10 pfu events with the four different PPS input groups and computing the resulting HSS and TSS scores, KWL17 considered an alternative situation in which the forecasted smaller (1 pfu < Ip < 10-pfu) events are also counted as hits. This increased the number a and decreased b of their contingency tables by the same number of small events, with resulting increases in the skill scores. This was equivalent to validating PPS against a smaller defined event threshold, but they left c and d unchanged in each contingency table before recalculating TSS and HSS. With a lower threshold, however, both a and c will increase and b and d decrease by the same numbers. Properly accounting for accompanying changes in c and d may have produced decreases of their associated skill scores, rather than the increases they reported. One can find an optimum forecast model performance by varying the threshold size, but this results in a model which may well be unsatisfactory for the model user. Here we seek an alternative basis for deciding on the forecast value of a model and in particular whether there is a way to evaluate a model for space weather forecasting based on the calculated importance of observed events. We want to make these new evaluations within the context of skill scores using the 2 × 2 contingency Table 1. Despite its flaws detailed above, the table has four categories readily interpretable in terms of forecasts and observations and serves as input into the well defined standard skill scores.

Table 2 lists the 138 KWL17 SEP events consisting of 67 ≥ 10 pfu and 71 small (1 pfu < Ip < 10 pfu) observed SEP events and their associated

Table 2
Continued

Peak	UT	Ip	> M5 flare	> M5 flare	8800 MHz
Date	Time	(p/cm^2 s sr)	Only	& 8800 MHz	> 500 sfu
5/9/1992	20:35	9.18	e	e	e
6/25/1992	22:35	36.6	a	a	a
10/31/1992	6:50	210	a	a	a
11/2/1992	5:50	312	a	a	a
3/4/1993	14:00	2.45	e	e	e
3/12/1993	20:10	4.71	b	b	e
9/25/1993	5:40	1.34	e	e	e
2/20/1994	6:00	1.51	e	e	e
9/3/1994	9:10	37	c	c	c
9/4/1994	9:05	16.3	c	c	c
9/5/1994	10:15	11.2	c	c	c
9/6/1994	9:45	58.3	c	c	c
11/4/1997	9:20	9.98	d	b	e
11/6/1997	18:25	115	a	a	a
4/21/1998	13:25	103	c	c	c
5/2/1998	15:20	24.3	c	a	a
5/6/1998	8:55	19.3	a	a	a
8/25/1998	1:50	10.9	c	c	c
9/30/1998	19:55	30.3	c	c	c
11/14/1998	9:50	27.9	c	c	c
6/2/1999	7:30	2.24	e	e	e
2/18/2000	10:15	1.26	e	e	e
6/10/2000	18:10	6.25	d	b	e
7/15/2000	9:35	1,670	c	a	a
7/22/2000	13:10	1.57	e	e	e
9/12/2000	22:50	1.95	e	e	e
10/16/2000	10:05	1.44	e	e	e
11/9/2000	3:40	1880	a	a	c
11/24/2000	18:00	4.98	e	b	e
11/26/2000	20:10	19	a	a	a
1/28/2001	21:05	1.89	e	e	e
4/3/2001	7:45	53.5	c	c	c
4/9/2001	19:45	1.2	d	b	e
4/10/2001	14:55	3.69	d	b	e
4/12/2001	17:50	5.75	b	b	e
4/15/2001	15:40	275	a	a	a
4/18/2001	6:40	40	c	c	c
5/20/2001	11:30	1.52	e	e	e
6/15/2001	17:30	1.7	d	e	e
8/16/2001	3:55	144	c	c	c

PPS forecasts. The first three columns give event dates, approximate peak times, and Ip. The last three columns give the event classifications of Table 1 using the three model inputs of KWL17. For each input all Ip ≥ 10 pfu events are either a, correctly forecasted (hits); or c, not forecasted (misses). Each small (1 pfu < Ip < 10 pfu) event is b, forecasted but not observed as an Ip ≥ 10-pfu event (a subset of all false alarms); d, forecasted as no event and not observed as a 10-pfu event (a subset of all correct null forecasts); or e, no forecast was made and not observed as a 10-pfu event. The last group, e, results from no PPS runs due to lack of suitable flare or radio burst events. It lies outside the context of Table 1 but merits consideration, as discussed below. Events are shown in Figures 1, 2, and 3, which give a synoptic view of the SEP peak-intensity power-law size distribution. Some of the X-ray fluence values of Figure 1 were computed from incorrect flare onset and end times (Swalwell et al., 2018), but that result is not relevant to our exploration of a new scoring methodology.

In this reevaluation of skill scores, we want to provide a weighting to events that measures the importance of each event to the user. In the case of SEP events, we assume that it is not event numbers, but the total amount of radiation in those events that the user wants accurately forecasted. The event fluence, the time-integrated particle intensity of interest, has been shown to scale with the event Ip (Kahler & Ling, 2018), so each event is weighted by its value of Ip as a measure of its relative importance. As evident in the plots, power-law size distributions may confer approximately equal weights to the few large events and the many small events. The goal is to maintain the format of Table 1 and the original value of n with new weighted events a', b', c', and d' rather than the original event numbers and then calculate new skill scores with those weighted events.

We first adhere to the original concept of including only observed events above the 10-pfu threshold. The relative sums of event intensities Ci of unforecasted events c (events in blue to the right of the dashed line in Figures 1–3) and Ai of the forecasted events a (subset events in yellow to the right of the dashed line in Figures 1–3) provide the basis of revised weighted values a' = (a + c) × Ai/(Ai + Ci) and c' = (a + c) × Ci/(Ai + Ci), keeping b and d fixed and a' + c' = a+c. The chief weakness of this scheme is that we can't assign intensities to elements b and d, consisting of unobserved events, so at this point, we do not change b or d. The original table event numbers a to d still provide guidance in terms of how much SEP intensity is observed and not observed, and n is kept constant, so a' + b + c' + d = a + b + c + d = n. Figure 4 shows a schematic contingency table with the intensity exchanges we discuss here indicated with the red letters and arrow.

We next deal with the second population of 71 small (1 pfu < Ip < 10 pfu) observed events, which exceeds the total number of 67 Ip ≥ 10 pfu events in our data set. They are shown in the two bins to the left of the dashed lines of Figures 1–3 and fall into three categories, the first being those forecasted by PPS as Ip > 10 pfu events, but observed as small events and shown in yellow in Figures 1–3, a subset of the number b in Figure 1, with total intensities Bi. The second category is events observed with a forecast of no event, a subset of group d of Table 1 with total intensities Di. The third category is events for which small or no X-ray flares or 8800 MHz bursts precluded any PPS forecast and are hence excluded from all groups

Table 2
Continued

Peak	UT	Ip	> M5 flare	> M5 flare	8800 MHz
Date	Time	(p/cm^2 s sr)	Only	& 8800 MHz	> 500 sfu
9/25/2001	22:30	273	c	a	a
10/2/2001	8:45	24.5	c	c	c
10/22/2001	21:05	2.5	d	e	e
11/6/2001	2:20	2,120	c	a	a
11/23/2001	12:25	162	a	a	a
12/26/2001	7:30	180	a	a	a
4/21/2002	10:25	208	a	a	a
8/22/2002	5:10	5.98	d	b	e
8/24/2002	2:50	76.2	a	a	a
11/9/2002	23:25	1.37	e	e	e
11/10/2002	5:40	1.46	e	e	e
5/31/2003	5:20	2.92	d	b	e
10/26/2003	19:30	10.3	a	c	a
10/27/2003	3:20	8	e	b	e
10/29/2003	1:00	1630	a	a	a
11/2/2003	22:10	153	a	a	a
11/5/2003	6:10	9.75	b	b	e
7/26/2004	22:50	1.86	e	e	e
9/19/2004	20:10	2.5	e	e	e
11/1/2004	7:00	5.64	e	e	e
11/7/2004	23:25	4.93	d	b	e
11/10/2004	10:20	13.2	a	a	a
1/16/2005	17:10	11	c	a	a
1/17/2005	17:00	350	a	a	a
1/20/2005	7:10	1070	a	a	a
6/16/2005	23:50	7.05	e	e	e
7/15/2005	4:50	1.75	e	b	e
8/23/2005	2:35	3.95	b	b	e
9/10/2005	2:50	50.6	c	c	c
9/14/2005	10:10	1.47	e	b	e
12/7/2006	18:30	103	c	c	c
12/13/2006	5:30	167	a	a	a
12/15/2006	0:15	11.6	a	a	a
6/7/2011	10:25	14.4	c	c	a
8/4/2011	8:05	8.43	d	b	e
8/9/2011	8:55	8.65	b	b	e
9/7/2011	5:00	1.61	d	b	e
1/23/2012	15:30	76.18	c	a	a
1/28/2012	1:00	47.2	a	a	a
3/7/2012	15:25	299.6	c	c	c

of Table 1. We let the total intensities of that group be Ei. The categories of Bi, Di, and Ei are shown in black letters in Figure 4. The basic question is how to reward or penalize the forecast model for forecasting or missing small events. We want a clear separation of all observed intensities into the a (forecasted) and c (missed) categories of Figure 4 while maintaining the a to d elements format of Table 1.

We treat the peak intensities of the three categories of small events in the following ad hoc way. Bi + Ai form the total forecasted SEP intensities. Initial weighted element a' is replaced with a" = a' × (Ai + Bi)/Ai to keep a" weighted by the ratio of the total forecasted intensities to the correctly forecasted intensities. To keep a new b' + a" the same normalized value of a' + b, we make b' = b − a' × (Bi/Ai). We treat the total intensities of elements c' and d similarly, except for the additional step of including intensities of the third category of unforecasted small events, defined above as Ei, in the new element c". Thus c" = c' × (Ci + Di + Ei)/Ci, and d' = d − c' × (Di + Ei)/Ci. While this procedure of dealing with small events is somewhat arbitrary, it does accomplish several goals. The sums of all forecasted intensities, Ai + Bi, are retained in the a" and b' elements, and a" + b' = a' + b. Similarly, the sums of all unforecasted intensities, Ci + Di + Ei are retained in the new c" and d' elements and c" + d' = c' + d. The Ei term is awkward because, being neither a forecasted nor an above threshold event, it was not contained in the original contingency matrix. Adding it to the total intensity of c" penalizes the model for its failure to forecast the associated small events composing Ei. In our intensity-weighted redistribution, a" + b' + c" + d' = a + b + c + d = n is conserved. Finally, we would emphasize that if only large events are of concern to the user, reevaluation of the contingency table for small events Bi, Di, and Ei may be ignored because their sum is 266.95 pfu, two orders of magnitude less than Ai + Ci = 28,265.35 pfu. That imbalance may not be the case for other applications, however.

In Table 3 we compare the original event number-based contingency elements of the three PPS forecast inputs of KWL17 with their revised values based on sums of event intensities. We compare the HSS and TSS scores along with the POD, which now measures the fraction of all SEP intensity forecasted by the model, and the FAR.

The intensity-based scores of the first group of ≥M5 flares are only slightly improved over the number-based group, probably because the forecasted (33) and missed (34) events are evenly divided and a" and c" are only slightly shifted from a and c. The results for both 8800-MHz groups are very different, however, showing large adjustments in a" and c", with corresponding large increases in HSS, TSS, and POD. The large number of false alarms b = 197 and 208, for those cases results in only slight declines in the intensity-based FAR. An initial assessment of the identical number-based a = 44 and b = 23 values of the two 8800-MHz models suggests little choice in their performance, but the intensity-based skill scores show that the 8800-MHz and > M5 flare input outperforms the 8800-MHz ≥ 500 sfu input.

Because of their small sizes, the 71 small (1 < Ip < 10 pfu) events of this study play only a minimal role in the intensity weightings above. However, we can ask how they compared with the many runs of the PPS that produced correctly forecasted null events, the group d of Table 1. Those d numbers were 603, 417, and 356 for the three models shown in Table 3.

Table 2
Continued

Peak	UT	Ip	> M5 flare	> M5 flare	8800 MHz
Date	Time	(p/cm^2 s sr)	Only	& 8800 MHz	> 500 sfu
3/13/2012	18:40	24.09	a	a	a
5/17/2012	3:00	78.29	c	a	a
7/7/2012	7:25	1.83	d	b	e
7/9/2012	1:30	3.31	d	b	e
7/19/2012	12:20	5.17	b	b	e
7/23/2012	22:15	3.24	e	e	e
4/11/2013	14:10	8.52	d	e	e
5/22/2013	22:30	26.4	c	c	c
9/30/2013	7:15	1.83	e	e	e
12/28/2013	22:15	1.56	e	e	e
1/6/2014	10:25	11.9	c	c	c
1/8/2014	6:00	48.09	c	a	a
2/20/2014	8:45	3.59	e	e	e
2/25/2014	21:05	3.27	d	e	e
2/28/2014	9:00	1.28	e	e	e
4/18/2014	14:55	2.44	d	b	e
9/3/2014	13:50	1.69	e	e	e
9/11/2014	4:25	4.47	d	b	e
10/29/2015	6:15	5.71	e	e	e

Note. a, b, c, and d elements are shown in Table 1 and Figure 4. Elements e are events for which no forecast was made and no > 10 pfu event was observed.

Abbreviations: PPS, proton prediction system; SEP, solar energetic particles.

The 71 small events might be expected to show up predominately in group b of false positives or of d, the large number of true negatives. However, we find corresponding totals of only 28, 27, and 3 of the 71 events in the b and d groups of Table 3. We note that KWL17 erroneously attributed 22 small events to group b for the input of 8800-MHz bursts > 500 sfu, where the correct number is 3, and 68 of the 71 small events were outside any forecasted group. Thus, in each model input, no forecasts were made for the majority of the small SEP events, so those events, with total intensities Ei (Figure 4) lie outside their associated contingency tables.

4. Extreme Intensity Variations with Fixed Event Numbers

We discussed in Section 1 the extreme condition in which the same skill scores result for two model inputs identical except that one forecasts a large event and misses a small one and the other input has the reverse forecast. As a demonstration of the effect of total intensity on our modified skill scores using our $E > 50$ MeV SEP events, we construct two extreme outcomes for each of our three inputs of Table 3 and Figures 1–3. We keep the same numbers of observed and missed events, a and c, given in Table 3, but from the 67 events of Ip > 10 pfu we first assume the worst case that a consists of the smallest Ip events and c the largest Ip events. Then we assume the reverse best case that a consists of the largest Ip events and c the smallest. Here we ignore the population of Ip < 10 pfu events, so the replacement of a and c numbers with their associated total intensities a′ and c′ varies only those parameters, not b and d. Table 4 compares the skill scores with the two extreme intensity configurations for each of the three inputs. Figures 5 and 6 contrast the intensity distributions for the worst and best cases of the example 8800-MHz and ≥ M5 flare input model of the PPS.

We find that the HSS and TSS scores turn negative in each of the worst cases, but stay positive for the best cases. POD values depend only on the varied a′ and c′ values, so are driven close to 0 and 1 for the worst and best cases. Variations of the FAR $= \text{a}'/(\text{a}' + \text{b})$ are limited by the large fixed values of b in each case. These contrived cases demonstrate that when space weather models are evaluated without consideration of the sizes or weightings of the individual events of the a and c elements of the contingency table, evaluation information is ignored.

5. Discussion

When the population of all space weather events of a particular type are of equal sizes or impacts, then only the numbers of forecasted and missed events are of consequence to the user. We note that Lopez et al. (2007) approached this ideal in testing various models forecasting whether the geosynchronous GOES-10 was inside or outside the magnetopause during a two-day disturbed period. They treated each of the total 2,880 min as individual events of equal weight in their contingency table to compare the forecasted and observed location of the GOES-10. However, most situations are defined by many small events and few large events, likely in the form of a power law (Aschwanden, 2019) or lognormal distribution (Verbeeck et al., 2019) for the entire population. For practical forecasting and user value, a size threshold is invoked, and this leads to the issue addressed here, of differentiating numbers of events from their total intensities or impacts. The standard skill scores, based solely on numbers of events, give some first-order guidance for evaluating forecast models, but the selected input variables of the models may well be biased toward selections of events of different sizes, and that is what we try to mitigate with our intensity-based skill scores.

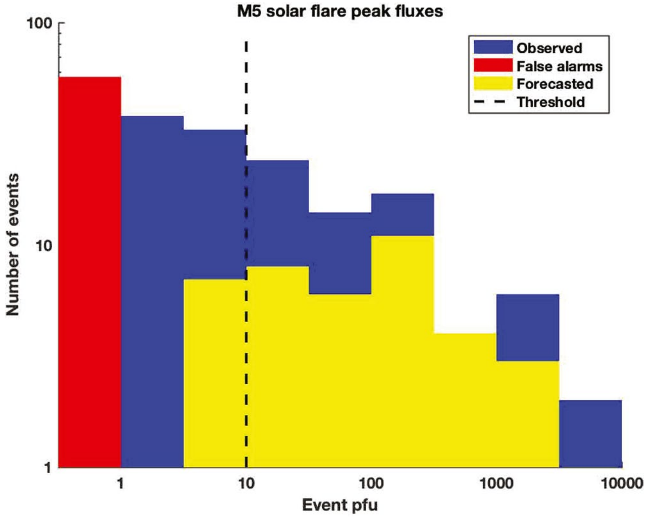

Figure 1. Size distribution of SEP events of KWL17 in units of pfu for the PPS model input of ≥ M5 peak flare fluxes. The pink bar on the left indicates the total number of forecasted false alarms of ≥ 10 pfu events with the ≥ M5 solar X-ray flare input option of the PPS. Numbers of observed events in each half decade >1 pfu are shown in blue. Dashed vertical line shows the 10-pfu threshold for the PPS forecast model. Yellow areas show the subset of events forecasted as ≥ 10 pfu events; those left of the dashed line are a subpopulation of the false positive group b and those to the right constitute all of group a. Blue events left of the dashed line are small events, which are not included in the number-based elements of Table 3 because no PPS forecast was made. PPS, proton prediction system; SEP, solar energetic particles.

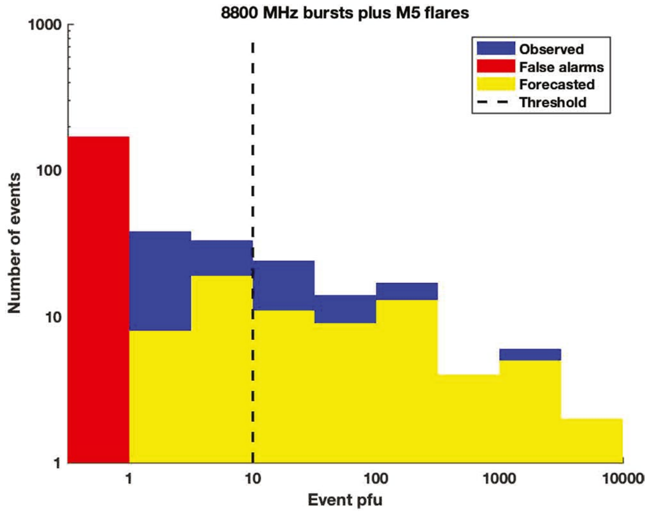

Figure 2. The same format as Figure 1 but for the PPS input option of 8800-MHz bursts with associated ≥ M5 solar X-ray flares. PPS, proton prediction system.

We have kept to the contingency format of Table 1, for which each element is well defined in terms of the observations and forecasts. This is advantageous for using the redefined event outcomes in the familiar skill score values and definitions. The primary drawback of our scheme of replacing numbers of events with their intensities is the asymmetry between known intensities of observed events (elements a and c) and undefined intensities of the false positive forecasts (element b). For the large SEP events, we elected to keep the redefined $a' + c' = a + c$, while retaining unchanged the values of b and d. For the small events, we reward the model forecasts by moving those event intensities Bi from the false negatives of b to a' (Figure 4) and reducing b by a comparable amount. Similarly, we have penalized the forecasts by moving observed intensities Di from the d element to the total intensities of missed events of c'. To account for the many small events for which no model forecast was done, we further penalize the models by adding those intensities Ei to element c' and decrease element d by the same amount. Better mathematical procedures may be possible, but each step has resulted in better or worse skill scores (Table 3) to reflect the inclusion of small events in the models.

The three PPS model inputs of KWL17 were reevaluated with traditional skill scores after taking account of both the >10 pfu and the 1 to 10 pfu populations of observed SEP events to account for intensities of all events, above and below the model thresholds. The revised skill scores (Table 3) better reflected the merits of the models, in particular, the difference between the two inputs based on 8800 MHz radio bursts, which appeared quite similar in their event number forecasts. We then carried out an exercise to look for the extreme limits of the inputs by assuming that forecasted events were only the worst-case smallest and then only the best-case largest of the same 67 targets SEP events. Skill score results (Table 4) were dramatically different, as expected.

We suggest that our intensity-based evaluation scheme could easily be applied to other forecast models using the binary outcomes of Table 1. The primary challenge would be to determine a reasonable method of weighting the forecast event variables, and that may mean incorporating a user impact in place of or combined with the size value of the predictand. One could expand beyond the simple fixed cost and loss (C and L) values of previous work (Mozer & Briggs, 2003; Owens & Riley, 2017; Park et al., 2017; Wilks, 2001) to assign economic values dependent on event intensities. The qualitative test of the method is then whether the chosen weighting scheme appears to perform better than the standard event number-based method at differentiating model outcomes, as we have shown in Table 3 for the PPS.

Perhaps more important is to seek and test alternative analytic methods to our ad hoc procedure indicated in Figure 4. We emphasize again that this work is only an exploration, whose robustness and utility remain undefined. The utility of POD and FAR and the sensitivity to bias B and event rate are also unknown. A comparison of Receiver Operating Characteristic (ROC; Wilks, 2006, chapter 7) diagrams could furnish a clear comparison of the merits of intensity-based versus number-based versions of a given model. The motivation here is that valuable observational intensity or size information is discarded when we reduce the data down to categorical event numbers to be employed in contingency tables (Table 1). The scheme here uses the previously discarded information while retaining the simplicity of scalar outputs of the standard HSS and TSS evaluation tools.

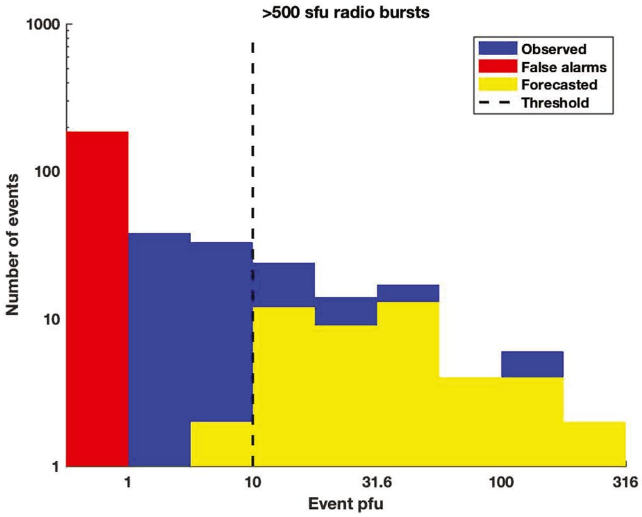

Figure 3. The same format as Figure 1 but for > 500 sfu 8800-MHz bursts.

A situation faced in terrestrial weather forecasting is an emphasis on extreme events, in which one selects for evaluating forecast models only forecasts of the very largest events with serious economic consequences. This selectivity, at the expense of overall forecasting merit, is the forecaster's dilemma, discussed in the Introduction. It can, however, lead to the discredit of models that perform well outside the extreme range and/or support for models that systematically overforecast. For PPS forecasts we propose here to weight events with intensities Ip, but an alternative method would be simply to raise the forecast threshold for unweighted events, say to 1,000 pfu, keeping to the deterministic forecast format of Table 1. An increased threshold might be equivalent or even preferred if the impacts of large events were sufficiently greater than those of sub-threshold events and an impact difference among the range of larger events were not significant. We have already seen in Section 3 that the 71 small (<10) SEP events of the PPS study had a minimal impact on the skill scores of models with intensity weightings. In situations with different event weighting schemes

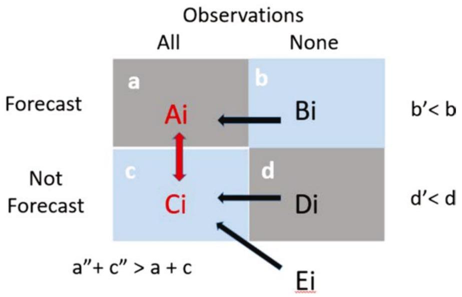

Figure 4. Contingency table showing original locations of SEP intensities. Sums of large events are Ai and Ci, in red font and sums of minor events are Bi, Di, and Ei. The first step (red arrow) is only to weight new elements a' and c' in proportion to their intensities Ai and Ci, with a' + c' (not shown) = a + c. In the next step (black arrows) the additional weighting of a' with Bi reflects the successful model forecasting of events with small intensities. Similarly, the weighting of c' with Di and Ei reflects the model failure to forecast events with those SEP intensities. The result is increased a” and c” and decreased b' and d', but a” + b' + c” + d' = a + b + c + d. SEP, solar energetic particles.

Table 3
Comparison of Verification Measures of PPS Forecasts Based on Event Numbers versus Total SEP Intensities Ip for Three Input Groups of KWL17

Solar variable	Forecast	Observed	Not observed	HSS	TSS	POD	FAR
Number-based	Yes	33	64	0.33	0.40	0.49	0.66
≥M5 flares	No	34	603				
Intensity-based	Yes	35.58	63.90	0.36	0.43	0.53	0.64
≥M5 flares	No	32.05	602.47				
Number-based	Yes	44	197	0.16	0.34	0.66	0.82
8800-MHz &M5	No	23	417				
Intensity-based	Yes	58.71	196.67	0.25	0.55	0.87	0.77
8800-MHz &M5	No	8.92	416.70				
Number-based	Yes	44	208	0.13	0.29	0.66	0.83
8800-MHz > 500	No	23	356				
Intensity-based	Yes	53.98	207.95	0.19	0.43	0.80	0.79
8800-MHz > 500	No	13.65	355.42				

Abbreviations: PPS, proton prediction system; SEP, solar energetic particles.

reflecting more equity of impact among events of various sizes, and with flatter parent size distributions, selecting a higher event-size threshold could on the other hand lead to degraded skill scores failing to characterize true model values for forecasting the smaller but still significant events. The variation of event weighting schemes and associated size thresholds provides a new parameter space for exploring alternative constructions of deterministic forecast models.

Table 4
Comparison of Verification Measures of PPS Forecasts Based on Extremes of Total SEP Intensities Ip for Three Input Groups of KWL17

Solar variable	Forecast	Observed	Not observed	HSS	TSS	POD	FAR
Lowest Ip	Yes	1.94	64	−0.67	−0.67	0.03	0.97
>M5 flares	No	65.06	603				
Highest Ip	Yes	65.06	64	0.62	0.88	0.97	0.50
>M5 flares	No	1.94	603				
Lowest Ip	Yes	4.60	197	−0.13	−0.25	0.07	0.98
8800-MHz & M5	No	62.40	417				
Highest Ip	Yes	62.40	197	0.27	0.61	0.93	0.76
8800-MHz & M5	No	4.60	417				
Lowest Ip	Yes	4.60	208	−0.15	−0.30	0.07	0.98
8800-MHz > 500	No	62.40	356				
Highest Ip	Yes	62.40	208	0.24	0.56	0.93	0.77
8800-MHz > 500	No	4.60	356				

Abbreviations: PPS, proton prediction system; SEP, solar energetic particles.

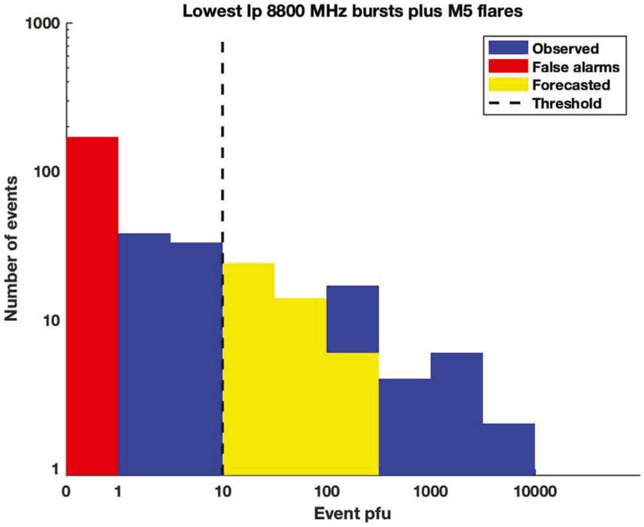

Figure 5. The same format as Figure 2 but for 8800-MHz bursts with the extreme example of the 44 forecasted events corresponding to the smallest of the 67 observed SEP events. SEP, solar energetic particles.

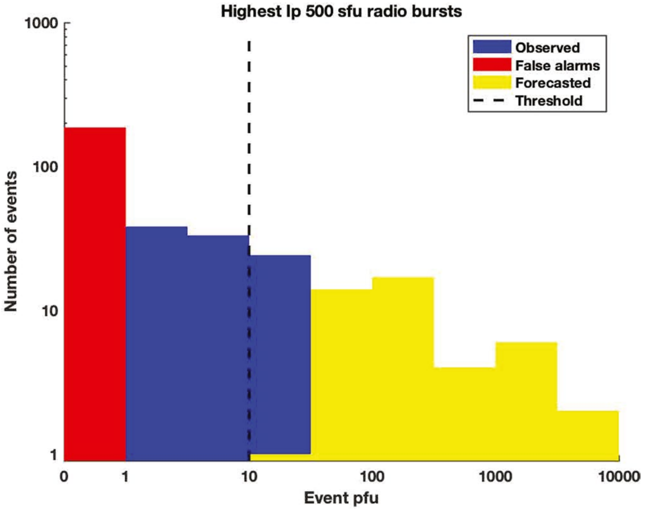

Figure 6. The same format as Figure 2 but for 8,800-MHz bursts with the extreme example of the 44 forecasted events corresponding to the largest of the 67 observed SEP events. SEP, solar energetic particles.

Data Availability Statement

All data and methodology for this paper are described in Sections 3 and 4 and are based on data of Table 2.

Acknowledgments

S. Kahler was funded by AFOSR Task 18RVCOR122 and benefitted from discussion with J. Mozer. H. Darsey was supported by the AFRL Phillips Scholars Program. The authors acknowledge the detailed criticisms and extensive suggestions of the editor and two reviewers of this work, which greatly changed the presentation.

References

Ahluwalia, H. S. (2019). Changes of space weather and space climate at Earth orbit: An update. *Advances in Space Research*, *64*, 1093–1099. https://doi.org/10.1016/j.asr.2019.05.046

Aschwanden, M. J. (2019). Self-organized criticality in solar and stellar flares: Are extreme events scale-free? *Astrophysical Journal*, *880*, 105. https://doi.org/10.3847/1538-4357/ab29f4

Barnes, G., Leka, K. D., Schrijver, C. J., Colak, T., Qahwaji, R., Ashamari, O. W., et al. (2016). A comparison of flare forecasting methods. I. Results from the "All-clear" workshop. *Astrophysical Journal*, *829*, 89. https://doi.org/10.3847/0004-637X/829/2/89

Belov, A., Kurt, V., Mavromichalaki, H., & Gerontidou, M. (2007). Peak-size distributions of proton fluxes and associated soft X-ray flares. *Solar Physics*, *246*, 457–470. https://doi.org/10.1007/s11207-007-9071-x

Bloomfield, D. S., Higgins, P. A., McAteer, R. T. J., & Gallagher, P. T. (2012). Toward reliable benchmarking of solar flare forecasting methods. *Astrophysical Journal Letters*, *747*, L41. https://doi.org/10.1088/2041-8205/747/2/L41

Bobra, M. G., & Couvidat, S. (2015). Solar flare prediction using SDO/HMI vector magnetic field data with a machine-learning algorithm. *Astrophysical Journal*, *798*, 135. https://doi.org/10.1088/0004-637X/798/2/135

Bobra, M. G., & Ilonidis, S. (2016). Predicting coronal mass ejections using machine learning methods. *Astrophysical Journal*, *821*, 127. https://doi.org/10.3847/0004-637X/821/2/127

Casati, B., Wilson, L. J., Stephenson, D. B., Nurmi, P., Ghelli, A., Pocernich, M., et al. (2008). Forecast verification: Current status and future directions. *Meteorological Applications*, *15*, 3–18. https://doi.org/10.1002/met.52

Chiarini, P. (2013). Space weather in the EU's FP7 space theme. *Journal of Space Weather and Space Climate*, *3*, E01. https://doi.org/10.1051/swsc/2013054

Cliver, E. W., & D'Huys, E. (2018). Size distributions of solar proton events and their associated soft X-Ray flares: Application of the maximum likelihood estimator. *Astrophysical Journal*, *864*, 48. https://doi.org/10.3847/1538-4357/Aad043

Echer, E., Gonzalez, W. D., & Tsurutani, B. T. (2011). Statistical studies of geomagnetic storms with peak Dst ≤ − 50nT from 1957 to 2008. *Journal of Atmospheric and Solar-Terrestrial Physics*, *73*, 1454–1459. https://doi.org/10.1016/j.jastp.2011.04.021

Falconer, D. A., Moore, R. L., Barghouty, A. F., & Khazanov, I. (2014). MAG4 versus alternative techniques for forecasting active region flare productivity. *Space Weather*, *12*, 306–317. https://doi.org/10.1002/2013SW001024

Gopalswamy, N. (2017). Extreme solar eruptions and their space weather consequences. In N. Buzulukova (Ed.), Extreme events in geospace Origins, predictability, and consequences, (pp. 37–63). Cambridge, MA: Elsevier (ISBN: 97800-12-812700-1).

Gopalswamy, N., Akiyama, S., Yashiro, S., Michalek, G., Xie, H., & Mäkelä, P. (2020). Effect of the weakened heliosphere in solar cycle 24 on the properties of coronal mass ejections. Journal of Physics: Conference Series, 1620(012005). Bristol, UK. March 9 - 13, 2020. IOP Publishing. https://doi.org/10.1088/1742-6596/1620/1/012005

Henley, E. M., & Pope, E. C. D. (2017). Cost-loss analysis of ensemble solar wind forecasting: Space weather use of terrestrial weather tools. *Space Weather*, *15*, 1562–1566. https://doi.org/10.1002/2017SW001758

Inceoglu, F., Jeppesen, J. H., Kongstad, P., Marcano, N. J. H., Jacobsen, R. H., & Karoff, C. (2018). Using machine learning methods to forecast if solar flares will be associated with CMEs and SEPs. *Astrophysical Journal*, *861*, 128. https://doi.org/10.3847/1538-4357/aac81e

Jackson, B. V., Yu, H. S., Buffington, A., Hick, P. P., Tokumaru, M., Fujiki, K., et al. (2019). A daily determination of BZ using the Russell-McPherron effect to forecast geomagnetic activity. *Space Weather*, *17*, 639–652. https://doi.org/10.1029/2018SW002098

Jiggens, P., Heynderickx, D., Sandberg, I., Truscott, P., Raukunen, O., & Vainio, R. (2018). Updated model of the solar energetic proton environment in space. *Journal of Space Weather and Space Climate*, *8*, A31. https://doi.org/10.1051/swsc/2018010

Kahler, S. W., Cliver, E. W., & Ling, A. G. (2007). Validating the proton prediction system (PPS). *Journal of Atmospheric and Solar-Terrestrial Physics*, *69*, 43–49. https://doi.org/10.1016/j.jastp.2006.06.009

Kahler, S. W., & Ling, A. G. (2018). Relating solar energetic particle event fluences to peak intensities. *Solar Physics*, *293*, 30. https://doi.org/10.1007/s11207-018-1249-x

Kahler, S. W., White, S. M., & Ling, A. G. (2017). Forecasting E > 50-MeV proton events with the proton prediction system (PPS). *Journal of Space Weather and Space Climate*, *7*, A27. https://doi.org/10.1051/swsc/2017025

Lerch, S., Thorarinsdottir, T. L., Ravazzolo, F., & Gneiting, T. (2017). Forecaster's dilemma: Extreme events and forecast evaluation. *Statistical Sciences*, *32*, 106. https://doi.org/10.1214/16-STS588

Lopez, R. E., Hernandez, S., Wiltberger, M., Huang, C.-L., Kepko, E. L., Spence, H., et al. (2007). Predicting magnetopause crossings at geosynchronous orbit during the Halloween storms. *Space Weather*, *5*(1), S01005. https://doi.org/10.1029/2006SW000222

Marsh, M. S., Dalla, S., Dierckxsens, M., Laitinen, T., & Crosby, N. B. (2015). SPARX: A modeling system for solar energetic particle radiation space weather forecasting. *Space Weather*, *13*, 386–394. https://doi.org/10.1002/2014SW001120

McIntosh, S. W., Chapman, S., Leamon, R. J., Egeland, R., & Watkins, N. W. (2020). Overlapping magnetic activity cycles and the sunspot number: Forecasting sunspot cycle 25 amplitude. *Solar Physics*, *295*, 163. https://doi.org/10.1007/s11207-020-01723-y

Mozer, J. B., & Briggs, W. M. (2003). Skill in real-time solar wind shock forecasts. *Journal of Geophysical Research*, *108*(A6), 1262. https://doi.org/10.1029/2003JA009827

Murphy, A. H. (1985). Decision making and the value of forecasts in a generalized model of the cost-loss ratio situation. *Monthly Weather Review*, *113*(3), 362–369. https://doi.org/10.1175/1520-0493(1985)113%3C0362:DMATVO%3

Núñez, M., Nieves-Chinchilla, T., & Pulkkinen, A. (2019). Predicting well-connected SEP events from observations of solar EUVs and energetic protons. *Journal of Space Weather and Space Climate*, *9*, A27. https://doi.org/10.1051/swsc/2019025

Owens, M. J., & Riley, P. (2017). Probabilistic solar wind forecasting using large ensembles of near-sun conditions with a simple one-dimensional "upwind" scheme. *Space Weather*, *15*, 1461–1474. https://doi.org/10.1002/2017SW001679

Papaioannou, A., Anastasiadis, A., Kouloumvakos, A., Paassilta, M., Vainio, R., Valtonen, E., et al. (2018). Nowcasting solar energetic particle events using principal component analysis. *Solar Physics*, *293*, 100. https://doi.org/10.1007/s11207-018-1320-7

Park, J., Moon, Y.-J., Choi, S., Baek, J.-H., Cho, K.-S., & Lee, K. (2017). Application of decision-making to a solar flare forecast in the cost-loss ratio situation. *Space Weather*, *15*, 704–712. https://doi.org/10.1002/2016SW001532

Richardson, I. G., Mays, M. L., & Thompson, B. J. (2018). Prediction of solar energetic particle event peak proton intensity using a simple algorithm based on CME speed and direction and observations of associated solar phenomena. *Space Weather*, *16*, 1862–1881. https://doi.org/10.1029/2018SW002032

Ryan, D. F., Dominique, M., Seaton, D., Stegen, K., & White, A. (2016). Effects of flare definitions on the statistics of derived flare distributions. *Astronomy & Astrophysics*, *592*, A133. https://doi.org/10.1051/0004-6361/201628130

Schrijver, C. J., Dobbins, R., Murtagh, W., & Petrinec, S. M. (2014). Assessing the impact of space weather on the electric power grid based on insurance claims for industrial electrical equipment. *Space Weather*, *12*, 487–498. https://doi.org/10.1002/2014SW001066

Schrijver, C. J., Kauristie, K., Aylward, A. D., Denardini, C. M., Gibson, S. E., Glover, A., et al. (2015). Understanding space weather to shield society: A global road map for 2015-2025 commissioned by COSPAR and ILWS. *Advances in Space Research*, *55*, 2745–2807. https://doi.org/10.1016/j.asr.2015.03.023

Schrijver, C. J., & Siscoe, G. L. (2010). *Heliophysics: Space Storms and Radiation: Causes and Effects*. Cambridge, MA: Cambridge University Press.

Sillmann, J., Thorarinsdottir, T., Keenlyside, N., Schaller, N., Alexander, L. V., Hegerl, G., et al. (2017). Understanding, modeling and predicting weather and climate extremes: Challenges and opportunities. *Weather and Climate Extremes*, *18*, 65–74. https://doi.org/10.1016/j.wace.2017.10.003

Stephenson, D. B. (2000). Use of the "odds ratio" for diagnosing forecast skill. *Weather Forecasting*, *15*, 221–232. https://doi.org/10.1175/1520-0434(2000)015<0221:UOTORF>2.0.CO;2

Svalgaard, L. (2020), Prediction of Solar Cycle *25*, arXiv:2010.02370 [astro-ph.SR].

Swalwell, B., Dalla, S., Kahler, S., White, S. M., Ling, A., Viereck, R., & Veronig, A. (2018). The reported durations of GOES soft X-ray flares in different solar cycles. *Space Weather*, *16*, 660–666. https://doi.org/10.1029/2018SW001886

Thomson, A. W. P. (2000). Evaluating space weather forecasts of geomagnetic activity from a user perspective. *Geophysical Research Letters*, *27*, 4049–4052. https://doi.org/10.1029/2000GL011908

Tobiska, W. K., Atwell, W., Beck, P., Benton, E., Copeland, K., Dyer, C., et al. (2015). Advances in atmospheric radiation measurements and modeling needed to improve air safety. *Space Weather*, *13*, 202–210. https://doi.org/10.1002/2015SW001169

Verbeeck, C., Kraaikamp, E., Ryan, D. F., & Podladchikova, O. (2019). Solar flare distributions: Lognormal instead of power law? *Astrophysical Journal*, *884*, 50. https://doi.org/10.3847/1538-4357/ab3425

Verbeke, C., Mays, M. L., Temmer, M., Bingham, S., Steenburgh, R., Dumbović, M., et al. (2019). Benchmarking CME arrival time and impact: Progress on metadata, metrics, and events. *Space Weather*, *17*, 6–26. https://doi.org/10.1029/2018SW002046

Wilks, D. S. (2001). A skill score based on economic value for probability forecasts. *Meteorological Applications*, *8*, 209–219. https://doi.org/10.1017/S1350482701002092

Wilks, D. S. (2006). *Statistical methods in the Atmospheric Sciences* (Vol. 91). Elsevier.International Geophysical Series

Zheng, Y., Ganushkina, N. Y., Jiggens, P., Jun, I., Meier, M., Minow, J. I., et al. (2019). Space radiation and plasma effects on satellites and aviation: Quantities and metrics for tracking performance of space weather environment models. *Space Weather*, *17*, 1384–1403. https://doi.org/10.1029/2018SW002042

Zhong, Q., Wang, J., Meng, X., Liu, S., & Gong, J. (2019). Prediction model for solar energetic proton events: Analysis and verification. *Space Weather*, *17*, 709–726. https://doi.org/10.1029/2018SW001915