Updated verification of the Space Weather Prediction Center's solar energetic particle prediction model

Christopher C. Balch[1]

Received 15 May 2007; revised 24 August 2007; accepted 10 September 2007; published 12 January 2008.

[1] This paper evaluates the performance of an operational proton prediction model currently being used at NOAA's Space Weather Prediction Center. The evaluation is based on proton events that occurred between 1986 and 2004. Parameters for the associated solar events determine a set of necessary conditions, which are used to construct a set of control events. Model output is calculated for these events and performance of the model is evaluated using standard verification measures. For probability forecasts we evaluate the accuracy, reliability, and resolution and display these results using a standard attributes diagram. We identify conditions for which the model is systematically inaccurate. The probability forecasts are also evaluated for categorical forecast performance measures. We find an optimal probability and we calculate the false alarm rate and probability of detection at this probability. We also show results for peak flux and rise time predictions. These findings provide an objective basis for measuring future improvements.

Citation: Balch, C. C. (2008), Updated verification of the Space Weather Prediction Center's solar energetic particle prediction model, *Space Weather*, *6*, S01001, doi:10.1029/2007SW000337.

1. Introduction

[2] The production of energetic particles by solar activity was first discovered over 60 years ago [*Forbush*, 1946]. In addition to the intrinsic scientific interest in these solar energetic particles (SEP), there are a number of practical applications for prediction of SEPs because of their impact on today's modern, technologically driven society. A primary example is manned spaceflight, which requires an understanding of the radiation hazards posed by SEPs to astronauts [*Cucinotta et al.*, 2002]. Similar concerns are also a subject of study for the newly emerging enterprise of space tourism [*Collins*, 2006] as well as for high-flying and commercial airlines [*Beck et al.*, 2005; *Dryer et al.*, 2005; *Getley et al.*, 2005]. SEPs are also known to affect spacecraft operations due to the interaction of high-energy particles with spacecraft electronics, which can lead to data errors and even unexpected behavior of the vehicle [*Iucci et al.*, 2005; *Dyer et al.*, 2004; *Feynman and Gabriel*, 2000]. In addition, energetic particles are able to reach the heights of 60–90 km in the polar ionosphere (the *D* region) and create a layer of charge that partially or completely absorbs high-frequency (HF) radio waves, thereby affecting long-distance radio communication and radar systems that operate in the HF band [*Hargreaves*, 2005; *Hunsucker*, 1992]. Because of these effects there is a societal need to specify and predict the energetic particle environment in order to understand operational difficulties and to mitigate these impacts.

[3] The requirement for predictions leads naturally to questions about cause and effect and also raises issues concerning what kinds of observable phenomena can provide information about the likelihood for occurrence or nonoccurrence of SEPs. With regard to physical causes there are three key components: the particle source population, the acceleration process, and the transport of the accelerated particles to the affected system. Each of these components poses a significant challenge to predictions. For example, *Desai et al.* [2006] studied the abundance characteristics of 64 SEP events in the 0.1–10 MeV/nucleon energy range and concluded that the seed population is highly variable and includes significant contributions from superthermal particles through which an interplanetary shock passes. Analyses of the particle acceleration process make frequent use of the diffusive shock acceleration (DSA) mechanism in the vicinity of shocks driven by coronal mass ejections (e.g., see *Zank et al.* [2005] and also *Roussev et al.* [2004]). The theory of diffusive shock acceleration, however, is still a work in progress. For example, *Sokolov et al.* [2006] argue that the spectrum of energetic particles accelerated by DSA depends on the magnetic

[1] NOAA Space Weather Prediction Center, Boulder, Colorado, USA.

field geometry of the shock because of a turbulent feedback mechanism in quasi-parallel shocks. A further complication arises because some SEP events appear to have a contribution due to stochastic acceleration from resonant wave-particle interactions in solar flares [*Miller and Reames*, 1996; *Temerin and Roth*, 1992]. In some cases, both of these mechanisms may be at work and affect the overall event profile and particle abundances [e.g., see *Cane et al.*, 2006].

[4] In addition to the problems of source population and particle acceleration, the transport of particles from the acceleration region to the affected system is also a challenging problem in its own right (recently reviewed by *McKibben* [2005]). For example, *Reames* [1999] argues for nearly scatter-free transport for gradual SEP events. However, *Dalla et al.* [2003] argue that Ulysses observations of several large SEP events provide evidence for an important role for cross-field diffusion (see also *Giacalone and Jokipii* [2001]). In addition, *Pei et al.* [2006] demonstrate complications that arise in determining particle transport when considering a more realistic configuration for the interplanetary magnetic field by constructing a Parker spiral that superposes a stochastically varying component.

[5] Because such complex physical processes are still an area of active research, utilization of physics-based numerical models has not yet evolved to the point where they can be used for forecasting. There have been some efforts in this direction. For example, *Aran et al.* [2006] have constructed a code that uses a large database of SEP event profiles and then provides a predicted, interpolated SEP profile using the heliolongitude and initial CME shock speed as input parameters. However, the model assumes that the observation of a CME with a shock is a sufficient condition for an SEP event and therefore does not address the important issue of SEP probability given the observed conditions. Other physics-based model development efforts and their limitations were recently reviewed by *Lario* [2005].

[6] Because of these difficulties, current methods for forecasting SEPs necessarily rely on observations of associated precursor phenomena. Such prediction schemes are intrinsically not deterministic but rely on empirical, statistical relationships between characteristics of observed phenomena and SEP parameters. Several examples of this approach have been presented [*Gerontidou et al.*, 2006; *Belov et al.*, 2005; *Garcia*, 2004a, 2004b; *Kubo and Akioka*, 2004; *Gabriel and Patrick*, 2003; *Balch*, 1999; *Smart and Shea*, 1989]. Although these approaches do not involve predictions that are derived from first-principles physical laws (the inductive approach), they should not be considered "unphysical" but should be considered to be deductive approaches that use data to identify patterns or relationships among the parameters which in turn point to the underlying physical processes behind SEP generation. In this paper we present results from a comprehensive evaluation of the performance of the operational model currently used in operations at NOAA's Space Weather Prediction Center (SWPC) of the U.S. National Centers for Environmental Prediction (NCEP). This model was previously described by *Balch* [1999].

[7] To carry out this analysis for the SWPC proton prediction model, the existing list of SEP events maintained at SWPC (http://swpc.noaa.gov/ftpdir/indices/SPE.txt, based on 5-min averaged proton flux from the National Oceanic and Atmospheric Administration (NOAA) Geostationary Operational Environmental Satellites (GOES)) was extensively reviewed, improved, updated, and extended. In addition, flare associations were reviewed and improved and a complementary list of control flares was compiled (i.e., flares met the necessary conditions for an SEP but were not associated with an SEP). A revision of the list as a result of this study is currently planned but not complete as of publication submission date.

[8] The evaluation of the model necessarily introduces the topic of forecast verification. Since many in the space weather community may be primarily space scientists and may be less familiar with the "weather" part of space weather, we provide an introduction to some of the ideas and methodologies which are used to measure the quality of the forecasts produced by the model. These concepts have been well developed in the context of tropospheric weather forecasting over the last several decades and are readily adapted to space weather applications. In this paper we use these techniques to evaluate the model's probability forecasts. We also provide a summary of the prediction performance for maximum flux and for proton event rise time. We find that the probability predictions have quadratic score of 0.0250 (rms error of 0.158). We also find specific conditions where the model is systematically inaccurate or is unable to discriminate between proton events and control events. An optimal strategy is devised for using the model to make categorical forecasts, but even at the optimum we find that the false alarm rate is 55% and the probability of detection is 57%. The results show that there remains significant room for improvement. It is hoped that by clearly establishing the current baseline for performance of an existing model, it will be possible to objectively measure the improvement of newer, updated prediction models. We also plan to use the updated database to carry out a new statistical analysis for empirical SEP predictions. This new analysis will look carefully at solar precursors to find the best discriminators between activity associated with SEPs and activity not associated with SEPs. The results of this work will be submitted in a future publication.

2. Description of the Model

[9] The SWPC proton prediction model is based on the association of solar flares with SEP events. The SWPC defines an SEP event to be an enhancement of protons with energy $\geq$10 MeV in excess of 10 protons cm^{-2} s^{-1} ster^{-1} as measured by the GOES satellites at geosynchronous orbit. Input parameters for the prediction model are time-integrated soft X-ray flux, peak soft X-ray flux, the

occurrence or nonoccurrence of metric radio type II and type IV sweeps, and the location of the associated flare. The X-ray event parameters are derived from 1-min averages of soft X-ray flux using the GOES X-ray Sensor (XRS) (see *Garcia* [1994] for a description of XRS). Metric radio sweep events are reported routinely in real-time from the USAF Solar Electro-Optical Observatory Network (SEON). Solar flare location is derived either from ground-based solar observatories using hydrogen-alpha telescopes or from spaced-based instruments such as the GOES-12 Solar X-ray Imager (SXI) or the SOHO Extreme Ultraviolet Imaging Telescope (EIT). On the basis of the statistical information described by *Balch* [1999], the model provides forecasters with a probability for a proton event, a prediction for the maximum flux at 10 MeV, and the time of maximum of the proton event.

[10] Note that although no direct parameter characterizing a CME is used, there remains nonetheless a relationship between these observations and the characteristics of an associated CME. For example, in a study of a large number of CMEs using SOLWIND data, *Sheeley et al.* [1983] showed a clear relationship between the duration of an X-ray event and the likelihood for a CME. Furthermore, the association between SEPs and type II as well as type IV radio sweeps was recognized very early [*Wild et al.*, 1963]. In fact the radio signatures were interpreted as an indication of a shock wave propagating through the corona. (See *Cliver* [2000] for a historical review of the evolution of ideas about SEPs and associated electromagnetic phenomena).

[11] The probability prediction is based on three input parameters: the XRS event maximum, the time-integrated XRS flux, and the occurrence or nonoccurrence of type II or type IV radio sweeps. The time period for integration of the XRS flux starts with the onset of the X-ray event and ends with the "half-power" point during the decay phase of the X-ray flux. The half-power point is defined as that time after maximum when the flux decreases to a level halfway between the maximum flux and the preevent background level. The current model in operational use is based on the event data described by *Balch* [1999] which consisted of 88 proton events and 1334 control events. The events were classified into groups according to five possible ranges of integrated flux, five possible ranges of maximum flux, and four possible values for the radio sweep observations. This constituted a 5 × 5 × 4, three-dimensional discrete parameter space into which proton events and control events were subdivided. Within each point in parameter space, the number of proton events divided by the total number of events was calculated to estimate the probability for proton event occurrence. For points in parameter space where the sample size was too small (taken to be less than 10 events), one dimension of parameter space was removed and the probability was reestimated using two parameters. In addition, for those cases where a point in two-dimensional parameter space still had insufficient sample size, another dimension was removed and the probability was evaluated based on a single parameter. The specifics of these calculations are shown explicitly in Tables 6, 7, and 8 of *Balch* [1999].

[12] The prediction of the maximum flux of a proton event at 10 MeV is based primarily on a statistical relationship between the log of the peak flux of the proton events and the log of the integrated X-ray flux of the associated X-ray event. The prediction also includes information about the integrated X-ray flux of the most recent previous event that occurred in the same active region (if applicable) as this was found to provide a slightly higher correlation between the predicted values and the observed values. The prediction formula is

$$\Phi_{P10} = 10\alpha_{PF} \times \left[\frac{X_{int}}{0.00987}\right]^{0.82}, \tag{1}$$

where Φ_{P10} is the predicted maximum flux at > 10 MeV,

$$\begin{aligned} \alpha_{PF} &= \left(\frac{X_{\mathrm{pfint}}}{0.167}\right)^{1.146} \quad \text{if } X_{\mathrm{pfint}} > 0.08, \\ \alpha_{PF} &= 1 \quad \text{if } X_{\mathrm{pfint}} \leq 0.08, \end{aligned} \tag{2}$$

X_{pfint} is the integrated X-ray flux of the previous event, and X_{int} is the integrated X-ray flux of the associated X-ray event. The correlation coefficient between the log of the predicted flux and the log of the observed flux was found to be 0.489 in the *Balch* [1999] study.

[13] The prediction of the rise time (i.e., the time difference between the X-ray event maximum and the proton event maximum) was based on an empirical relationship that was found with the location of the associated flare on the solar disk. The formula derived is

$$t_{rise} = t_{\min} + \left[\frac{longitude - l_{SE}}{\lambda}\right]^2, \tag{3}$$

where l_{SE} is the optimal sub-Earth longitude (78 degrees west), λ is a longitudinal scaling factor (18.1 degrees), and $t_{\min}$ is the minimum rise time (9.4 h). The model reflected the trend of the observations (see Figure 13 of *Balch* [1999]) but there was still a significant scatter of points about the prediction model. The standard error for the proton events in the 1999 study was found to be 22.5 h.

3. Description of the New Event Database

[14] Since the time of the 1999 study, solar cycle 23 has passed through solar maximum (April 2000) and has produced numerous proton events. This provides an opportunity to reexamine the model performance by the inclusion of these new events. The original 1999 paper covered events from 1976 to 1995: in this study, we consider two solar cycles of proton events as recorded from 1986 to 2004. The greater availability of digital data from GOES as well as ground-based optical and radio observatories has enabled a careful examination and considerable improvement in the quality of the event database as compared to what was previously possible.

[15] A review of all proton events from 1986 to 2004 was carried out by an examination of corrected, archived 5-min GOES particle flux. Because of the ready availability of the 5-min flux values it was possible to verify and improve existing proton event records. The result was a compilation of 165 events from 1986 to 2004. Each event includes several parameters derived from the GOES Energetic Particle Sensor (EPS) data, including onset time, threshold time (flux exceeding 10 particles cm^{-2} s^{-1} ster^{-1}), time of maximum, end time, maximum flux at $\geq$10 MeV, $\geq$30 MeV, $\geq$60 MeV, and $\geq$100 MeV, and event fluences (time integrated flux) for the same energies.

[16] A comparison of the corrected, 5-min GOES integrated particle flux with the existing list of 10 MeV proton events (http://swpc.noaa.gov/ftpdir/indices/SPE.txt) revealed inconsistencies between the peak fluxes of the events prior to 1990. The source of these discrepancies was the introduction of corrections to the calculation of integrated particle flux in January 1990 as described by *Onsager et al.* [1996] (see also http://goes.ngdc.noaa.gov/data/avg/readme.txt). These corrections were applied to the older GOES EPS data sets after the fact and resulted in corrections to the proton event data for the period prior to 1990. In addition to the corrected events, some events were removed from the original list since the corrected peak flux no longer exceeded the threshold of 10 particles cm^{-2} s^{-1} ster^{-1}. SWPC plans to apply these corrections to the event list on the Web site but this has not yet been completed as of the submission date of this publication.

[17] An additional modification of the SWPC proton event list was necessary for this analysis. Some of the proton events consisted of multiple injections of energetic particles from multiple solar events. Since the emphasis of this study is the prediction of energetic particles using solar observations, it was necessary to consider each injection to be a distinct, solar-generated proton event. This necessarily introduces some uncertainty in the onset times and end times for these special cases. This particular definition of a proton event will not be incorporated into the SWPC Web site event list because those events are defined strictly in terms of threshold crossings of the GOES $\geq$ 10 MeV proton flux.

[18] The association of GOES XRS events with the proton events was also reexamined. Readily available 1-min XRS digital data helped facilitate this process, and it was possible to improve the associations and the XRS event parameters relative to what had been done previously. In particular, the time-integrated x-ray flux was recalculated to insure that the parameter was derived consistently over the study period. The parameters derived from the XRS data set include begin time, maximum time, postmaximum half-power time, end time, peak flux, time-integrated flux, and background subtracted integrated flux.

[19] The association of ground-based H-alpha flare reports and radio sweep events was also reexamined. In addition to the event reports received from the USAF SEON network, it was possible to supplement the observations with reports from other observatories as archived by the National Geophysical Data Center. The ground based data quality was further enhanced by information about times of active observing. Thus associated activity could be classified according to three possibilities: (1) an associated event did occur, (2) an associated event did not occur, or (3) it is unknown whether an associated event occurred. The inclusion of this supplemental data together with the SEON report archive enabled the addition of optical location, type II occurrence, and type IV occurrence for a large majority of the events.

[20] In examining each of these events, it was found that 28 (17%) originated from sources behind the solar limb. Most of these originated behind the west limb, but two cases were found to have originated behind the east limb. Obviously, these events present a special kind of space weather challenge since some or all of the soft X-ray emission and radio emission may have been unobservable due to obscuration of the solar disk. The solar signatures needed for the prediction model were either unavailable or showed indications that they were affected by partial disk obscuration. Therefore for this analysis we will treat these as a priori missed events. If relevant observations could be made for activity behind the limb using a strategically placed spacecraft-based observatory, then we would expect the number of missed events to decrease.

[21] We also found that 9 (5.4%) of the SEP events attained threshold due to energetic storm particles (ESP), which are trapped particles that are observed in situ during passage of an interplanetary shock [*Cohen et al.*, 2005]. These events would not have reached threshold level if there were not an ESP component. These cases also present a special space weather prediction problem, since the physical processes depend primarily on properties of the shock and the nature of the ambient medium through which the shock passes, and much less directly on the characteristics of the associated solar event. Again, we will treat these as a priori missed events with respect to verification of the prediction model.

[22] Another special case was identified where two solar sources combined to produce an SEP event. Analysis of the timing of the solar activity and the proton flux profiles showed that neither solar source could have independently produced an SEP event but that the SEP event was a consequence of two proton enhancements occurring in close time coincidence. Therefore these solar sources were not considered to be proton producing events individually and the SEP event was discarded from the analysis.

[23] For the remaining 127 events, analysis of the event parameters showed that all were associated with an XRS event with peak flux in the 1–8 Å band of greater than 2.44×10^{-6} Watts m^{-2}, integrated X-ray flux greater than 9.93×10^{-3} Joules m^{-2}, and background-subtracted integrated X-ray flux greater than 5.95×10^{-3} Joules m^{-2}. Thus we consider these to be necessary conditions for a proton event, with the qualification that this does not include events generated from backside solar activity or the small

ESP events we mentioned previously. We note that since we have 38 discarded events it will not be possible to detect more than 127 out of 165 events. This places an upper limit on the probability of detection at 77%.

[24] In order to validate the probability prediction, we supplemented the proton event data with a complete set of XRS events which meet these necessary conditions but did not produce a proton event (i.e., the control event database). These data were derived in a similar manner as described for the proton events. The GOES XRS 1-min data for the same time period were carefully examined for events meeting the criteria. In addition, archives of XRS events from NGDC and SWPC were used to supplement the analysis. For each XRS event, radio sweep (type II and type IV) information and optical flare information were derived from the same sources as mentioned previously. The final result was a list of 3656 events from 1986 to 2004 which meet the necessary conditions but were not associated with a proton event. Of these, 20 events were associated with an enhancement of 10 MeV proton flux, but the flux did not reach event level criteria. These borderline cases are retained for the verification analysis, but they will be removed in future work when we analyze the data to discriminate between proton producing events and control events.

4. An Overview of Forecast Verification Measures

[25] In this section we describe well-developed concepts for forecast verification that are used in the context of tropospheric weather forecasting. The interested reader may want to refer to *Jolliffe and Stephenson* [2003] and references therein for an in-depth review of this topic.

[26] Forecast verification seeks to evaluate the skill or value of a forecast. Generally speaking, there are three classes of users who need verification information, and the interpretation of verification results will vary somewhat with the needs of each user class. The first of these are users who need to make operational or economic decisions based on the forecast. In this case the user needs to know the distribution of observations given the prediction and can use this information to make a risk assessment to determine an optimal operational strategy that will minimize losses. A second class of user has an administrative or programmatic perspective; in this case the objective is to track whether forecasts are getting better or worse. For this user class the key is to be able to judge the forecast skill in relative terms, that is, measure whether forecast system A is better than forecast system B. Verification also provides a means for an administrator to measure the impact on forecasting skill due to a particular data stream, a particular model, or a particular forecaster training program and can guide prioritization of future resource allocations to meet space weather objectives. A third class consists of users who seek to improve scientific understanding through forecast verification. In particular, verification analysis may help these users identify conditions where the model is doing poorly and may help focus future model development efforts or observational programs on leveraged topics where increased understanding will result in major improvements to forecast performance.

[27] The accuracy of a forecasting system is defined to be ''the average degree of correspondence between individual forecasts and observations'' [*Murphy and Daan*, 1985]. The most common approach for measuring the accuracy of probability forecasts, a key output of the SWPC model proton prediction model, is to compare the prediction vector $\boldsymbol{f}$ (which contains probabilities scaled between 0 and 1) with the corresponding observation vector $\boldsymbol{o}$ (where o_i = 0 if no event occurs but o_i = 1 if an event does occur), and to calculate the mean square error between the predictions and the observations:

$$QR = (1/N)\sum_{i=1}^{N}\left(f_i - o_i\right)^2, \tag{4}$$

where N is the number of model predictions.

[28] The quantity QR is known as the quadratic score or the Brier score [*Brier*, 1950]. Note that smaller Brier scores indicate better accuracy and that a perfect Brier score of 0.0 would require a 0.0 probability forecast every time an event did not occur and a 1.0 probability forecast every time an event did occur.

[29] The accuracy can be compared with a reference forecast to calculate a measure of relative skill. Typically, the observed rate of occurrence of the events relative to the total number of forecasts is used as a baseline. So, for example, we can calculate the occurrence rate for SEP events, $\overline{\mathrm{o}}$ = number of proton events/number of forecast runs. We then can calculate a reference quadratic score, QR*, based on a constant forecast of $\overline{\mathrm{o}}$. The skill score for the prediction model relative to this climatological forecast can be calculated as follows:

$$SS = (QR^* - QR)/QR^* \tag{5}$$

We can see from the formula that SS will approach zero as QR approaches QR^* (no skill), and SS will approach one as QR approaches zero (perfect skill).

[30] In order to understand the forecast system performance in greater depth, however, it is necessary to consider the joint probability distribution of the observations and forecasts [*Murphy and Winkler*, 1987]. Since for probability forecasts it is not possible to draw meaningful conclusions from the comparison of an individual forecast with an individual observation, it is necessary to use a statistical approach, so that we compare the correspondence between forecast probabilities and the observed relative frequency of the event. (Reviews for verification of probability forecasts may be found in the work of *Toth et al.* [2003], *Hsu and Murphy* [1986], and *Murphy and Daan* [1985]). So, for example, if we want to measure the performance of the forecasts when a given probability is predicted, say 60%, then we need a sufficient sample of such forecasts from which we can calculate the proportion

Table 1. Two-By-Two Contingency Table for Analyzing the Quality of Categorical Forecasts

	Event Observed		
Event Forecast	Yes	No	
Yes	A	B	nw
No	C	D	nn
	np	nc	N

of occurrences of the event to see how close the occurrence frequency is to 60%. If the predicted probability and the relative frequency occurrence are nearly the same, the forecasts are statistically consistent with the observations; if there is a significant difference between these quantities, however, we can say that the forecasts are not statistically consistent with the observations and that these forecasts have a bias (i.e., the forecasting system is underforecasting or overforecasting). This measure of consistency between the forecasts and observations is commonly referred to as forecast reliability, a term originally introduced by *Murphy* [1973]. It should be noted here that if the biases of a forecast system are well known, it is straightforward to apply bias corrections which will improve model reliability.

[31] However, statistical reliability by itself does not provide complete information about the usefulness of the probability forecasts. For example, if we use the statistical occurrence rate for an event over a period of time (e.g., 1 year during solar maximum), as a probabilistic forecast, we would expect the event occurrence frequency to be consistent with the forecast probability and hence for these forecasts to be highly reliable. However, this type of prediction (a climatology forecast) does not provide any ability to distinguish conditions when the event will occur with higher or lower frequency than the climatological rate. The degree to which a forecasting system can distinguish these circumstances is a measure of forecast resolution [*Hsu and Murphy*, 1986; *Murphy*, 1973]. For a perfectly resolved forecast system, the predictions divide the sample into two or more subsamples such that all forecasts $f_i < \overline{o}$ would have an event occurrence rate of 0.0 and all forecasts $f_i > \overline{o}$ would have an event occurrence rate of 1.0. Unlike reliability, it is not possible to apply the results of verification analysis to make improvements in resolution. Instead, more fundamental changes are needed in the observations, models, or forecaster training in order to better discriminate conditions which produce events from conditions which do not produce events.

[32] The analysis of *Murphy* [1973] and *Hsu and Murphy* [1986] shows that there is a close relationship between accuracy, reliability, and resolution, which can be expressed as follows:

$$QR = QR^* + REL - RES. \tag{6}$$

In this equation, *REL* is the reliability and *RES* is the resolution, defined by the following formulae:

$$REL = (1/N)\sum\nolimits_{i=1}^{T} N_i\left(\langle f_i\rangle - \langle o_i\rangle\right)^2, \tag{7}$$

$$RES = (1/N)\sum\nolimits_{i=1}^{T} N_i(\langle o_i\rangle - \overline{o})^2, \tag{8}$$

where T is the number of probability ranges. For each forecast probability range, N_i is the number of model runs, $\langle f_i\rangle$ is the mean forecast probability, and $\langle o_i\rangle$ is the event occurrence frequency. Smaller values for *REL* indicate better reliability and lower the overall mean square error. Larger values for *RES* indicate better resolution and also lower the overall mean square error. Therefore by analyzing reliability and resolution of probability forecasts, it is possible to identify model deficiencies and identify pathways for future improvement in overall accuracy. The specific application of these concepts will be demonstrated in the results section of the paper.

[33] The performance of probability forecasts can also be analyzed using tools that are applied to verify categorical (i.e., yes/no) forecasts. For categorical forecasts it is possible to evaluate performance in terms of false alarm rate (FAR), probability of detection (POD), and other measures we will define presently. In order to apply categorical measures to probability forecasts, it is necessary to define a probability threshold, p_t. We then consider prediction performance given that a warning is issued whenever the forecast probability is greater than or equal to p_t and such that no warning is issued if the forecast probability is less than p_t. In this context we can analyze the forecasts and observations in terms of a 2 × 2 contingency table, shown in Table 1, where we define the following variables: *A* is the number of hits, *B* is the number of false alarms, *C* is the number of missed events, *D* is the number of correct nulls, *N* is the total number of forecast model runs. The row totals are *nw*, number of warnings issued, and *nn*, number of cases for which a warning was not issued. The column totals are *np*, the number of proton events, and *nc*, the number of control events. The following statistical measures provide information about the quality of these categorical forecasts:

$$\begin{aligned}&\text{Probability of detection } (POD) = A/(A+C)\\&\text{False alarm rate } (FAR) = B/(A+B)\\&\text{Percent correct } (PC) = (A+D)/N.\end{aligned} \tag{9}$$

[34] It is possible to adjust the *Percent Correct* statistic by subtracting the number of forecasts that we would expect to be correct by chance. This provides a skill corrected verification measure known as the Heidke Skill Score (*HSS*) [*Heidke*, 1926] (see also *Wilks* [1995]). The number

of correct predictions by chance is derived according to the following argument:

$$\begin{aligned}&\text{Probability}(\text{event} = \text{Yes}) = (A + C)/N,\\&\text{Probability}(\text{forecast} = \text{Yes}) = (A + B)/N,\end{aligned} \tag{10}$$

and therefore the probability for a chance hit is

$$\text{P}(\text{event} = \text{Yes and forecast} = \text{Yes}) = (A + C)^*(A + B)/N^2. \tag{11}$$

The derivation for the probability of a chance correct null is derived as follows:

$$\begin{aligned}&\text{Probability}(\text{event} = \text{No}) = (B + D)/N,\\&\text{Probability}(\text{forecast} = \text{No}) = (C + D)/N,\end{aligned} \tag{12}$$

hence

$$\text{P}(\text{event} = \text{No and forecast} = \text{No}) = (B + D)^*(C + D)/N^2. \tag{13}$$

The combined probability for a chance correct forecast (hits and correct nulls) is

$$[(A + B)^*(A + C) + (B + D)^*(C + D)]/N^2. \tag{14}$$

We therefore derive the number of correct forecasts by chance to be

$$E = [(A + B)^*(A + C) + (B + D)^*(C + D)]/N \tag{15}$$

This leads to the definition of the Heidke Skill Score:

$$HSS = (A + D - E)/(N - E). \tag{16}$$

[35] The Heidke skill score can range from −1 for no correct forecasts (under certain conditions) up to +1 for all correct forecasts. A value of zero is interpreted to mean that the predictions that are no better than chance.

[36] For probability forecasts, we can consider the probability threshold, p_t, to be an independent variable that varies from 0.0 to 1.0 (or equivalently 0% to 100%). For each fixed value of p_t, the categorical statistics may be calculated; therefore we can consider the categorical quality measures *POD, FAR,* and *HSS* to be functions of the probability threshold p_t. Typically, the false alarm rate will decrease as the threshold level is increased. However, this is usually at the expense of a decrease in the probability of detection. In general we will be able to find an optimal skill score (using *HSS*) for some value of the probability threshold.

5. Verification Results

[37] With the event data, it was straightforward to test the model for performance during the past two solar cycles. We begin the discussion by looking at the results for the probability forecast.

[38] On the basis of the description of the model from *Balch* [1999] the probability was calculated for 3783 events, of which 127 were proton events and 3656 were control events. The events were grouped according to T = 11 ranges of predicted probabilities (see Table 2). On the basis of this grouping, we determine the accuracy, reliability, resolution, and skill of the prediction model and display this information using an attributes diagram as described by *Hsu and Murphy* [1986] (see Figure 1). Shown along the abscissa are values for average predicted probabilities for each group and shown along the ordinate are the corresponding observed occurrence rates. The numbers labeling each point indicate the sample size for each group, and the error bar is measure of uncertainty in the estimated proportion, based on the size of the sample. The REL = 0 dashed line indicates perfect reliability, the RES = 0 line indicates a forecasting system with no resolution (i.e., a constant forecast using the climatological occurrence rate).

[39] Accuracy of the probability forecasts was measured using the quadratic score, *QR*, which was found to be 0.0250. Since QR is simply the mean square error, the rms error is $\text{QR}^{1/2}$ = 0.158. This result is consistent with the 1999 paper where we found an rms error of 0.16.

[40] The accuracy was compared with a reference forecast based on the average occurrence rate of SEPs. The average occurrence rate (number of proton events/number of forecast runs) was found to be $\overline{\text{o}}$ = 127/3783 = 0.0336. A constant forecast of $\overline{\text{o}}$ has a quadratic score QR* = 0.0324; therefore the sample skill score is calculated to be

$$SS = (QR^* - QR)/QR^* = 0.230. \tag{17}$$

We note for purposes of comparison that *Hsu and Murphy* [1986] calculate a skill score of 0.32 for a sample of experimental precipitation probability forecasts with threshold amount of 0.01 inches and a skill score of 0.02 for similar forecasts for 0.25 inches.

[41] The terms from which reliability and resolution are calculated are also shown in Table 2. For each forecast probability range, N_i is the number of model runs, $\langle f_i \rangle$ is the mean forecast probability, and $\langle o_i \rangle$ is the event occurrence frequency. Note that $QR \rightarrow QR^*$ as $REL \rightarrow RES$, so we deduce that the skill score is zero when $REL = RES$. Thus the line in Figure 1 midway between REL = 0 and RES = 0 represent forecasts with no skill and is labeled in the figure with SS_i = 0. The verification curve lies above this line and shows positive skill for each of the probability ranges.

[42] Referring to Figure 1, we can assess the reliability of the model by comparison with the diagonally dashed line defined by $\langle \text{f}_\text{i} \rangle = \langle \text{o}_\text{i} \rangle$. As discussed previously, this dashed line defines perfect reliability where *REL* = 0. Points above the line indicate underforecasting and

Table 2. Attributes Table for Space Weather Prediction Center's Proton Prediction Model[a]

Forecast Probability	N_{tot}	N_{sep}	N_{ctrl}	$\langle f_i \rangle$	$\langle o_i \rangle$	REL_i ($\times 10^{-2}$)	RES_i ($\times 10^{-2}$)	SS_i	QR_i
0.00 – 0.05	3475	38	3437	0.005	0.011	0.004	0.05	0.01	0.01
0.05 – 0.15	38	4	34	0.088	0.105	0.029	0.51	0.15	0.09
0.15 – 0.25	109	13	96	0.184	0.119	0.418	0.73	0.10	0.11
0.25 – 0.35	28	15	13	0.300	0.536	5.556	25.21	6.06	0.30
0.35 – 0.45	44	16	28	0.379	0.364	0.024	10.89	3.35	0.23
0.45 – 0.55	40	17	23	0.457	0.425	0.099	15.32	4.69	0.25
0.55 – 0.65	49	24	25	0.596	0.490	1.126	20.81	6.07	0.26
0.65 – 0.75	0	0	0	-	-	-	-	-	-
0.75 – 0.85	0	0	0	-	-	-	-	-	-
0.85 – 0.95	0	0	0	-	-	-	-	-	-
0.95 – 1.00	0	0	0	-	-	-	-	-	-
All	3783	127	3656	0.030	0.034	0.07	0.82	0.230	0.0250

[a] Here 127 proton events and 3656 control events were used to produce 3783 model runs. Predictions from the model were grouped into probability ranges as indicated by the first column. The second, third, and fourth columns provide a count of the total number of events, the number of proton events, and the number of control events that fall into each probability range. Here $\langle f_i \rangle$ is the average probability forecast each interval and $\langle o_i \rangle$ is the occurrence frequency of proton events, N_{sep}/N_{tot}, for each interval. REL_i is the measure of reliability for each interval, defined as $(\langle f_i \rangle - \langle o_i \rangle)^2$, and the overall reliability is the weighted average of the REL_i using the number of events in each interval. Smaller values indicate higher reliability. The RES_i is a measure of resolution for each interval, defined as $(\langle o_i \rangle - \bar{o})^2$, where $\bar{o}$ is the overall occurrence frequency, N_{sep}/N_{tot} = 127/3783 = 0.034. Higher values indicate better resolution. SS_i is the sample skill score for each interval and QR_i is the quadratic score (Brier score) for each interval. Higher values of SS_i indicate better skill, and lower values of QR_i indicate better accuracy. The skill is calculated using QR_i relative to the quadratic score that results using the occurrence frequency, o_i. The overall skill score and quadratic scores are weighted averages over all of the intervals. The overall rms error is $\sqrt{QR}$ = 0.158.

points below the line indicate overforecasting. The overall reliability is calculated to be 0.0007. Most of the points show fairly good reliability, but one point in the probability range 0.25 – 0.35 clearly shows a region of significant underforecasting. In addition, there appears to be a tendency toward overforecasting in the 0.55 – 0.65 probability range. The reliability for each range of predicted probability is shown in Table 2 and confirms, quantitatively, the domains of relatively less reliable forecasts.

[43] An analysis of the underforecast events with predicted probabilities in the 25 – 35% led to the identification of a point in parameter space for which the original model statistics were significantly different than what was observed during the 1986 – 2004 interval. This point in parameter space consisted of events with integrated flux greater than 0.895 J m^{-2}, X-ray maximum greater than or

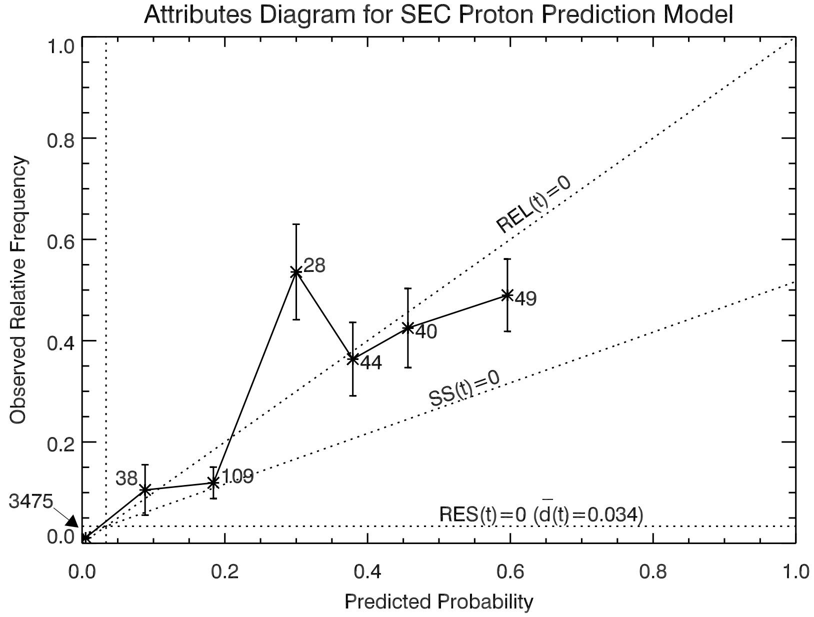

Figure 1. Attributes diagram for Space Weather Prediction Center's (SWPC) proton prediction model, based on events observed between 1986 and 2004.

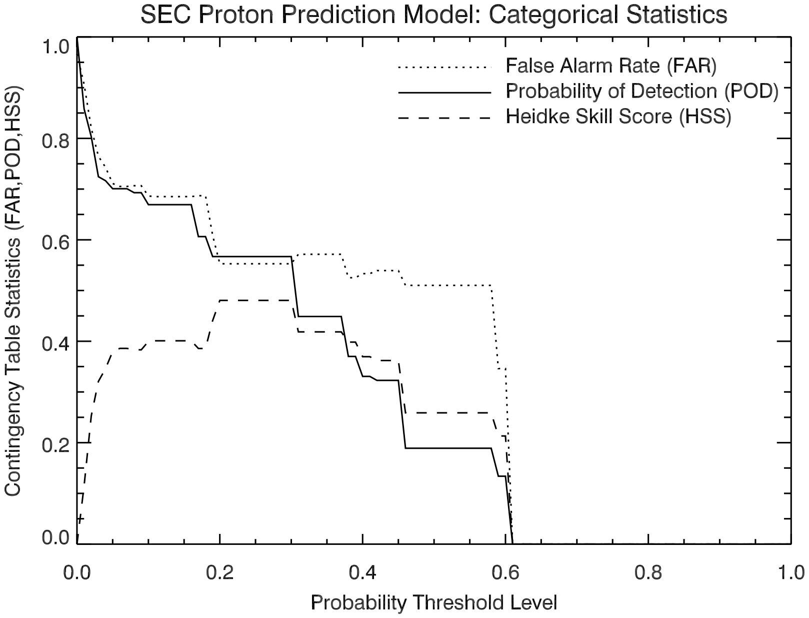

Figure 2. Categorical performance statistics for the proton prediction model as a function of probability thresholds.

equal to X7, and association of type II and type IV radio bursts. In the 1999 study, 3 out of 10 such events produced an SEP, suggesting a probability of 30%. However, the event data from 1986 to 2004 show that 10 out of 12 such events produced an SEP, suggesting a much higher probability of 83%. The discrepancy is an indication of a problem in using relatively small samples to estimate these probabilities. As was stated in the original 1999 paper, a 68% confidence interval for this category was 15.6–44.4%, but clearly this still leaves a 16% chance for a probability in excess of 44.4%. With the large number of proton event occurrence rates being made as a basis for the model, it certainly was likely that one or more of the estimated probabilities would be outside the 68% confidence interval.

[44] An analysis of the overforecast events with predicted probabilities in the 55–65% range shows a similar problem. In this case we find a point in parameter space (events with integrated flux 0.085–0.275 J m^{-2}, M3–M8, Type IV only) for which the SEP occurrence rate differs significantly. In the 1999 study the probability was estimated to be 58% (derived from 7 out of 12 such events), but for the 1986–2004 event data we find 7 out of 23 such events produce an SEP, an occurrence rate of 30%. Again we see that the relatively small sample size led to an inaccurate probability estimate that was incorporated in the model.

[45] The overall resolution of the forecasts is found to be 0.0082. An examination of the terms of the weighted sum for equation (8) shows that the smallest contributions to resolution (which detract from model accuracy) are from the 0.00--0.05, 0.05--0.15, and 0.15--0.25 probability ranges. This makes apparent a key weakness of the model: a significant fraction of the proton events fall in the lower probability ranges: for example, we see from Table 2 that 38/127 = 30% of the proton events are given a probability between 0.00 and 0.05. This indicates that more work is needed to find ways to distinguish the proton events from the control events under conditions for which the model produces such forecasts.

[46] This result is further strengthened by examining the contributions to the quadratic score from each of the probability ranges. Quadratic scores for each range are shown in Table 2, and the overall score *QR* is simply the weighted sum of the scores for each range, QR_i, where the weights are the fraction of events in each range. The contribution from the 0.00–0.05 range is (3475/3783) × 0.01 = 0.0092 which therefore contributes 0.0092/0.0250 = 37% of the mean square error.

[47] We consider next the performance categorical quality measures for probability of detection (*POD*), false alarm rate (*FAR*), and Heidke skill score (*HSS*), which, as discussed earlier, are functions of a probability decision threshold p_t. Using the proton prediction model and the data, these quality measures were calculated and are shown in Figure 2. An examination of the graph shows that the false alarm rate decreases as the threshold level is increased. As expected, we see that this decrease is at the expense of the probability of detection which also decreases with increasing threshold. We see that the optimal skill score (using *HSS*) is achieved for the range of probabilities from 20 to 30%. We also note that at the optimal point, the probability of detection is 57% (72/127), the false alarm rate is 55% (89/161), the percent correct is 96% (3639/3783), and the Heidke skill score is 0.48. This illustrates that even at the point of optimal skill, the

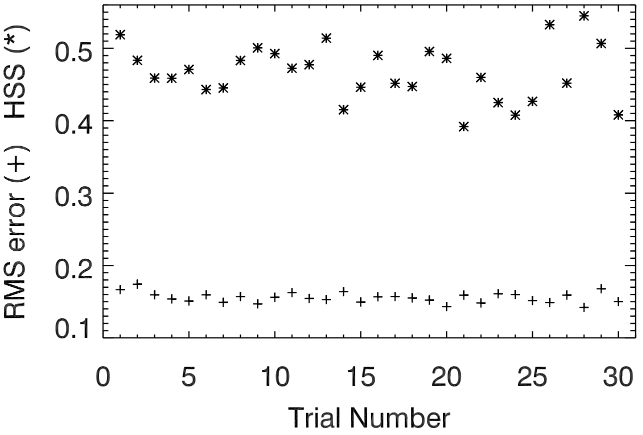

Figure 3. RMS error and Heidke skill scores for different versions of SWPC's proton prediction models. The structure of the input parameter space is the same for all of the versions, but for each version a different collection of proton events and control events has been used to derive the model probabilities. The mean rms error is found to be 0.156 ± 0.007, with the minimum value of 0.142 and a maximum value of 0.174. The mean HSS is found to be 0.47 ± 0.04, with a minimum value of 0.39 and a maximum of 0.54.

number of false alarms and missed events is significant, and there is clearly a need for improvement.

[48] A possible weakness in this analysis is that the model being verified was constructed from some of the events included in the verification data set. Proton events and controls from 1988 to 1997 were used in the earlier study and many of those events overlap with our 1986–2004 analysis interval. In order to address this issue, we divided the 1986–2004 events into two approximately equal groups. One group was used to construct new statistics for the model, using the same discrete structure for input parameter space, and the other group was used to verify the newly constructed model. The selection of events was done randomly. The process of event selection, model construction, and model verification was repeated for 30 independent trails. For each trail the rms error and the optimal Heidke skill score were calculated. Shown in Figure 3 are the results for rms error and optimal Heidke skill score for each of the trails. The mean rms error is found to be 0.156 ± 0.007, with the minimum value of 0.142 and a maximum value of 0.174. This provides a good estimate of the uncertainty in the rms error of 0.157 that was stated earlier. The mean value for HSS is found to be 0.47 ± 0.04, with a minimum value of 0.39 and a maximum of 0.54.

[49] Although it is beyond the scope of this paper, it should be pointed out that the HSS is only one alternative among various ways to optimize the utility of the forecasts. For some applications it may be more important not to miss any events than it is to experience occasional false alarms. On the other hand, a different application may be sufficiently affected by false alarms such that it is a better strategy to risk missing some of the events in order to lower the false alarm rate. The analysis can be quantified by comparing the cost of preventive action with the cost of a loss when an event occurs in the absence of preventive action. See *Weigel et al.* [2006] for a comprehensive discussion.

[50] The model predicted peak flux at $\geq$10 MeV for the 127 SEP events was compared with the observed peak flux and the result is shown in Figure 4. There continues to be a weak correlation between the logarithms of the predicted and observed peak flux: the current data show a correlation coefficient of 0.524 (c.f. with 0.489 from the previous study). We also find a significant scatter of the points about the perfect prediction line: the standard error is found to be 0.870 (c.f. with 0.764 from the previous study), so most of the time the predictions are within an order of magnitude of the observations but not always. This is similar to the result that was shown in the previous study [*Balch*, 1999].

[51] For the same SEP events we show the scatter of rise times as a function of longitude in comparison to the model formula, see Figure 5. The standard error in rise times is found to be 12.6 h, somewhat smaller than the 18.9 h that was reported in the previous study. Figure 6 shows a comparison of the predicted rise times with the observed rise times. The plot does show a correlation between the two quantities, but clearly there is significant scatter as we would expect, given the size of the standard error.

6. Summary

[52] The development of understanding and physical models to improve prediction of SEP events is an area of active research. However, at this time none of these appear to be sufficiently developed to be applied to real-time space weather operations, and there are still significant gaps in the knowledge and understanding of these events. Thus today's working prediction models are empirical and are based on statistical analysis of the characteristics of events that are associated with SEPs. In this paper we have evaluated the performance of one such model, the empirical proton prediction model currently being used at NOAA's Space Weather Prediction Center.

[53] To carry out this evaluation, we constructed a database of 127 proton events and associated flare signatures from 1986 to 2004. An analysis of the event data led us to find the following necessary conditions for a proton event: peak soft X-ray flux greater than 2.44×10^{-6} W m^{-2}, integrated X-ray flux greater than 9.93×10^{-3} J m^{-2}, and background subtracted integrated X-ray flux greater than 5.95×10^{-3} J m^{-2}. These conditions were then applied to construct a list of X-ray events over the same time period, and it was found that there were 3656 X-ray events that met the necessary conditions but were not associated with a proton event. The proton event and control event data

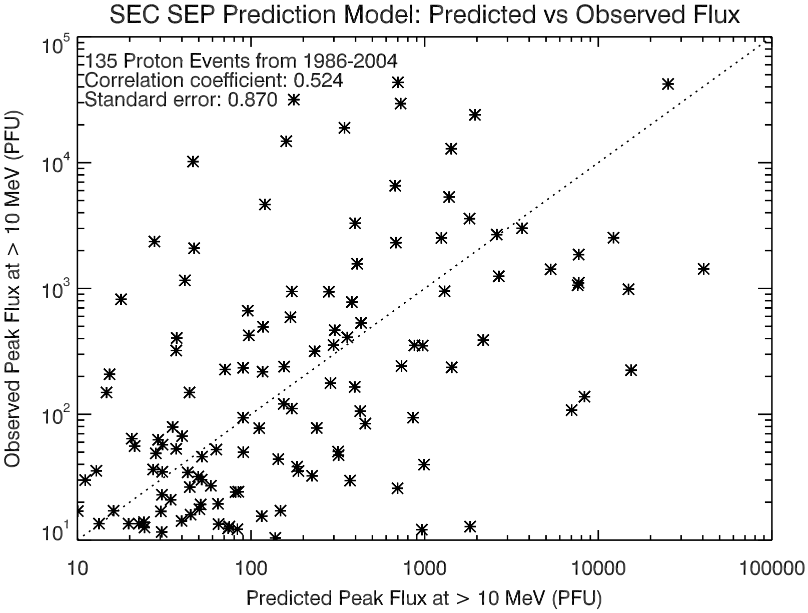

Figure 4. Comparison of predicted peak flux with observed peak flux for 127 SEP events from 1986 to 2004.

were used to verify the performance of the model's probability predictions. We find that the accuracy (mean square error) is 0.0250; hence the rms error is 0.158. The skill score relative to sample climatology is found to be 0.230. The accuracy statistic is shown to be composed of three components: climatology (contributes 0.0324), reliability (contributes 0.0007), and the resolution (contributes −0.0082). These results are depicted graphically with an attributes diagram. We find two conditions where the model probability is systematically inaccurate (poor reliability). Further analysis shows that the inaccuracy resulted from small sample sizes that were originally used to construct the model probabilities for these conditions. An additional problem is found concerning the resolution of the model. It is shown that 92% (3475/3783) of the model runs result in a probability prediction of 0–5%, of which ~1% end up as proton events. Unfortunately, this also means that the model predicts this probability range for 38 of the 127 proton events (~30%) SEP events.

[54] The probability results are also considered in the context of categorical forecast performance measures. The optimal Heidke skill score is achieved using a probability threshold in the 20–30% range and has a value of 0.48 ± 0.04.We find that at this optimal probability threshold the false alarm rate is 55% and the probability of detection is 57%.

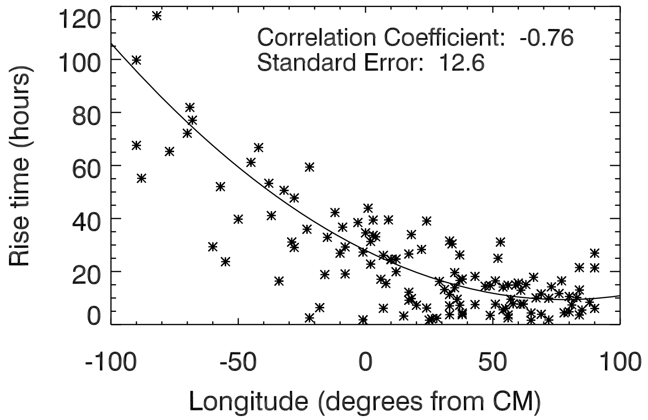

Figure 5. A plot of event rise times as a function of longitude. The smooth curve is the rise time prediction model.

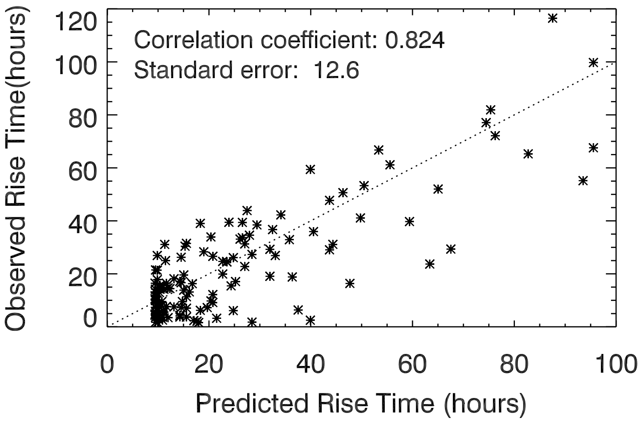

Figure 6. Comparison of predicted rise time with observed rise time.

[55] We also show results of peak flux prediction and rise time prediction. Both quantities show a correlation between the predicted and observed values but they also show considerable scatter and room for improvement.

[56] It should also be remembered that the above results do not include 28 SEP events that originated from activity behind the solar limb nor do they include nine events that were the result of energetic storm particles (ESP) only. Thus the current operational techniques necessarily suffer from lower probability of detection than described above. It may be possible to address the problem of behind-limb events by making use of observations from new, strategically placed spacecraft in the heliosphere (e.g., STEREO). Concerning the ESP events we note that they tend to have softer spectra and relatively small fluxes and may be negligible for some applications. Nonetheless, their occurrence does lower the overall SEP probability of detection and there is a need to develop predictive techniques for these particular events.

[57] Now that the database has been improved, extended, and updated, we have the opportunity to develop improved probability models. It is hoped that by including additional observational input such as CME speed, direction, and size it will be possible to reduce the current inaccuracies and improve the usefulness of the predictions. In addition, it is now possible to investigate other probabilistic questions, for example, probability for events at higher flux levels or at different energies. We can also look at probabilities for a variety of fluence levels at various energies.

[58] There continues to be a need to provide better predictions for peak flux, onset time, rise time, event duration, and spectral properties of SEPs. All of these need to be studied to discern how effective an empirical approach can be. Ultimately, space weather forecasters look forward to the day when physics-based models can surpass the limitations of what can be done empirically.

[59] **Acknowledgments.** The author thanks NOAA's National Geophysical Data Center, especially Ed Erwin and Helen Coffey, for their help in making the flare, radio, and GOES data archives complete. I also would like to thank Kent Doggett for helping me to understand some of the verification measures used in this paper. Also thanks to Joe Kunches for doing the initial work on assembling the proton event listing. I also gratefully acknowledge the help of Ed Cliver for his help and insight into the association of SEP events with various types of solar activity. Last, I gratefully acknowledge the illustrious career and life of Gary Heckman who was responsible for developing the original version of the proton prediction model at SWPC.

References

Aran, A., B. Sanahuja, and D. Lario (2006), SOLPENCO: A solar particle engineering code, *Adv. Space Res.*, *37*, 1240–1246.

Balch, C. C. (1999), SEC proton prediction model: Verification and analysis, *Radiat. Meas.*, *30*, 231–250.

Beck, P., M. Latocha, S. Rollet, and G. Stehno (2005), TEPC reference measurements at aircraft altitudes during a solar storm, *Adv. Space Res.*, *36*, 1627–1633.

Belov, A., H. Garcia, V. Kurt, H. Mavromichalaki, and M. Gerontidou (2005), Proton enhancements and their relation to the X-ray flares during the three last solar cycles, *Sol. Phys.*, *229*, 135–159.

Brier, G. W. (1950), Verification of forecasts expressed in terms of probability, *Mon. Weather Rev.*, *78*, 1–3.

Cane, H. V., R. A. Mewaldt, C. M. S. Cohen, and T. T. von Rosenvinge (2006), Role of flares and shocks in determining solar energetic particle abundances, *J. Geophys. Res.*, *111*, A06S90, doi:10.1029/2005JA011071.

Cliver, E. W. (2000), Solar flare photons and energetic particles in space, in *Acceleration and Transport of Energetic Particles Observed in the Heliosphere*, edited by R. A. Mewaldt et al., *AIP Conf. Proc.*, *528*, 21–31.

Cohen, C. M. S., E. C. Stone, R. A. Mewaldt, R. A. Leske, A. C. Cummings, G. M. Mason, M. I. Desai, T. T. von Rosenvinge, and M. E. Wiedenbeck (2005), Heavy ion abundances and spectra from the large solar energetic particle events of October–November 2003, *J. Geophys. Res.*, *110*, A09S16, doi:10.1029/2005JA011004.

Collins, P. (2006), Space tourism: From Earth orbit to the Moon, *Adv. Space Res.*, *37*, 116–122.

Cucinotta, F., G. Badhwar, P. Saganti, W. Schimmerling, J. Wilson, L. Peterson, and J. Dicello (2002), Space radiation cancer risk projections for exploration missions: Uncertainty reduction and mitigation, *NASA/TP 2002-210777*, NASA Johnson Space Cent., Houston, Tex.

Dalla, S., et al. (2003), Properties of high heliolatitude solar energetic particle events and constraints on models of acceleration and propagation, *Geophys. Res. Lett.*, *30*(19), 8035, doi:10.1029/2003GL017139.

Desai, M. I., G. M. Mason, R. E. Gold, S. M. Krimigis, C. M. S. Cohen, R. A. Mewaldt, J. E. Mazur, and J. R. Dwyer (2006), Heavy-ion elemental abundances in large solar energetic particle events and their implications for the seed population, *Astrophys. J.*, *649*, 470–489.

Dryer, C., F. Lei, A. Hands, S. Clucas, and B. Jones (2005), Measurements of the atmospheric radiation environment from CREAM and comparisons with models for quiet time and solar particle events, *IEEE Trans. Nucl. Sci.*, *52*(6), 2326–2331.

Dyer, C. S., K. Hunter, S. Clucas, and A. Campbell (2004), Observation of the solar particle events of October and November 2003 from CREDO and MPTB, *IEEE Trans. Nucl. Sci.*, *51*(6), 3388–3393.

Feynman, J., and S. B. Gabriel (2000), On space weather consequences and predictions, *J. Geophys. Res.*, *105*, 10,543–10,564.

Forbush, S. E. (1946), Three unusual cosmic ray increases possibly due to charged particles from the Sun, *Phys. Rev.*, *70*, 771–772.

Gabriel, S. B., and G. J. Patrick (2003), Solar energetic particle events: Phenomenology and prediction, *Space Sci. Rev.*, *107*, 55–62.

Garcia, H. A. (1994), Temperature and emission measure from GOES soft X-ray measurements, *Sol. Phys.*, *154*, 275–308.

Garcia, H. A. (2004a), Forecasting methods for occurrence and magnitude of proton storms with solar hard X rays, *Space Weather*, *2*, S06003, doi:10.1029/2003SW000035.

Garcia, H. A. (2004b), Forecasting methods for occurrence and magnitude of proton storms with solar soft X rays, *Space Weather*, *2*, S02002, doi:10.1029/2003SW000001.

Gerontidou, M. V., A. V. Belov, H. A. Garcia, H. A. Mavromichalaki, and V. G. Kurt (2006), Prospects of space weather prediction based on solar proton events, in *Recent Advances in Astronomy and Astrophysics, AIP Conf. Proc.*, *848*, 253–257.

Getley, I. L., M. L. Duldig, D. F. Smart, and M. A. Shea (2005), The applicability of model based aircraft radiation dose estimates, *Adv. Space Res.*, *36*, 1638–1644.

Giacalone, J., and J. R. Jokipii (2001), The transport of energetic particles and cosmic rays in the heliosphere, *Adv. Space Res.*, *27*, 461–469.

Hargreaves, J. K. (2005), A new method of studying the relation between ionization rates and radio-wave absorption in polar-cap absorption events, *Ann. Geophys.*, *23*(2), 359–369.

Heidke, P. (1926), Berechnung des Erfolges und der Gute der Windstarkevorhersagen im Sturmwarnungsdienst, *Geogr. Ann.*, *8*, 301–349.

Hsu, W., and A. H. Murphy (1986), The attributes diagram: A geometrical framework for assessing the quality of probability forecasts, *J. Int. Forecasting*, *2*, 285–293.

Hunsucker, R. D. (1992), Auroral and polar-cap ionospheric effects on radio propagation, *IEEE Trans. Antennas Propag.*, *40*(7), 818–828.

Iucci, N., et al. (2005), Space weather conditions and spacecraft anomalies in different orbits, *Space Weather*, *3*, S01001, doi:10.1029/2003SW000056.

Jolliffe, I. T., and D. B. Stephenson (2003), *Forecast Verification: A Practitioner's Guide in Atmospheric Science*, 240 pp., John Wiley, New York.

Kubo, Y., and M. Akioka (2004), Existence of thresholds in proton flares and application to solr energetic particle alerts, *Space Weather*, *2*, S01002, doi:10.1029/2003SW000022.

Lario, D. (2005), Advances in modeling gradual solar energetic particle events, *Adv. Space Res.*, *36*, 2279–2288.

McKibben, R. B. (2005), Cosmic-ray diffusion in the inner heliosphere, *Adv. Space Res.*, *35*, 518–531.

Miller, J. A., and D. V. Reames (1996), Heavy ion acceleration by cascading Alfvén waves in impulsive solar flares, in *High Energy Solar Physics*, edited by R. Ramaty, N. Mandzhavidze, and X.-M. Hua, *AIP Conf. Proc., 374*, 450–460.

Murphy, A. H. (1973), A new vector partition of the probability score, *J. Appl. Meteorol.*, *12*, 534–537.

Murphy, A. H., and H. Daan (1985), Forecast evaluation, in *Probability, Statistics, and Decision Making in the Atmospheric Sciences*, edited by A. H. Murphy and R. W. Katz, pp. 379–438, Westview, Boulder, Colo.

Murphy, A. H., and R. L. Winkler (1987), A general framework for forecast verification, *Mon. Weather Rev.*, *115*, 1330–1338.

Onsager, T. G., R. N. Grubb, J. Kunches, L. Matheson, D. M. Speich, R. D. Zwickl, and H. Sauer (1996), Operational uses of the GOES energetic particle detectors, in *GOES-8 and Beyond*, edited by E. R. Washwell, *SPIE Conf. Proc., 2812*, 281–290.

Pei, C., J. R. Jokipii, and J. Giacalone (2006), Effect of a random magnetic field on the onset times of solar particle events, *Astrophys. J.*, *641*, 1222–1226.

Reames, D. V. (1999), Particle acceleration at the Sun and in the heliosphere, *Space Sci. Rev.*, *90*, 413–491.

Roussev, I. I., I. V. Sokolov, T. G. Forbes, T. I. Gombosi, M. A. Lee, and J. I. Sakai (2004), A numerical model of a coronal mass ejection: Shock development with implications for the acceleration of GeV protons, *Astrophys. J. Lett.*, *605*, L73–L76.

Sheeley, N. R., Jr., R. A. Howard, M. J. Koomen, and D. J. Michels (1983), Associations between coronal mass ejections and soft x-ray events, *Astrophys. J.*, *272*, 349.

Smart, D. F., and M. A. Shea (1989), PPS-87 - A new event oriented solar proton prediction model, *Adv. Space Res.*, *9*, 281–284.

Sokolov, I. V., I. I. Roussev, L. A. Fisk, M. A. Lee, T. I. Gombosi, and J. I. Sakai (2006), Diffusive shock acceleration theory revisited, *Astrophys. J. Lett.*, *642*, L81–L84.

Temerin, M., and I. Roth (1992), The production of He-3 and heavy ion enrichment in He-3-rich flares by electromagnetic hydrogen cyclotron waves, *Astrophys. J. Lett.*, *391*, L105–L108.

Toth, Z., O. Talagrand, G. Candille, and Y. Zhu (2003), Probability and ensemble forecasts, in *Forecast Verification: A Practitioner's Guide in Atmospheric Sciences*, edited by I. T. Jolliffe and D. B. Stephenson, pp. 137–164, John Wiley, New York.

Weigel, R. S., T. Detman, E. J. Rigler, and D. N. Baker (2006), Decision theory and the analysis of rare event space weather forecasts, *Space Weather*, *4*, S05002, doi:10.1029/2005SW000157.

Wild, J. P., S. F. Smerd, and A. A. Weiss (1963), Solar bursts, *Annu. Rev. Astron. Astrophys.*, *1*, 291–366.

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

Zank, G. P., G. Li, G. M. Webb, J. A. Le Roux, V. Florinski, X. Ao, and W. K. M. Rice (2005), Particle acceleration at collisionless shocks: An overview, in *The Physics of Collisionless Shocks: 4th Annual IGPP International Astrophysics Conference*, edited by G. Li, G. P. Zank, and C. T. Russell, *AIP Conf. Proc.*, *78*, 170–179.

C. C. Balch, NOAA Space Weather Prediction Center, 325 Broadway, Boulder, CO 80305, USA. (christopher.balch@noaa.gov)