SPACE WEATHER, VOL. 4, S05002, doi:10.1029/2005SW000157, 2006

Decision theory and the analysis of rare event space weather forecasts

R. S. Weigel,[1] T. Detman,[2] E. J. Rigler,[3] and D. N. Baker[1]

Received 14 February 2005; revised 13 December 2005; accepted 16 December 2005; published 2 May 2006.

[1] Several basic results from decision theory as applied to rare event forecasts are reviewed, and an alternative method for comparing rare event forecasts is presented. A fundamental result is that for a large class of users only interested in economic utility, the relevant performance quantity is the number of correct and false alarm forecasts. This is contrasted with the reality that most forecast models are optimized to have a high data-model correlation, which does not always correspond to maximum economic utility. The value score (VS) developed by Wilks (2001) partially resolves this disconnect between modeler- and user-relevant metrics. Although the value score is closer to what is most likely of interest to a user, maximal VS does not necessarily correspond to maximal utility for the realistic case where the cost and benefit are dependent on the amplitude of the forecasted event. An alternative comparison and presentation method is proposed which may resolve this problem. For the class of users considered, full specification of model performance requires computation of the probability of correct, false alarm, and missed forecasts at several amplitude levels and warning time spans. Examples of the computations involved for the modeler and user are given for predictions of large-amplitude energetic electron fluence and geomagnetic storms parameterized by the *Dst* index.

Citation: 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.

1. Introduction

[2] Evaluation of a model's prediction performance in terms of a utility metric is a strong departure from the usual methods of model evaluation and parameter optimization, which are usually correlation based. That is, the parameters of a model are usually determined such that a quantity such as the mean square error between its prediction and a measured quantity is minimized. Moreover, evaluation of a model in terms of its ability to predict only events, as opposed to placing weight on every data point as is the case with correlation analysis methods, is also a departure from the norm. However, there is good reason to use correlation metrics instead of utility metrics. They are very general, and the computations involved in model optimization are straightforward for linear systems. The disadvantage of considering more user-relevant metrics is that a model cannot always be optimized for every possible user. In this work, we show that more user-relevant metrics can be defined while still maintaining generality.

[3] We summarize several mathematical aspects of evaluating a model on the basis of its ability to predict an event. In the space weather arena such analysis is fairly new, with recent evaluations including *Thomson* [2000], *Mozer and Briggs* [2003], and *Weigel et al.* [2003, 2004]. The motivation behind the application of decision theory to event forecasts is that the end user of a forecast must make a decision for action on the basis of each forecast [*Lindley*, 1985].

[4] Several recent works used decision theory related or motivated analysis on space weather problems. These analyses used various quantities as the measure of performance of the model in question. *Thomson* [2000] was the first to recognize the importance of decision theory in applications of space weather event predictions. The metrics considered were the likelihood ratio (LR) and the loss structure (K). *Mozer and Briggs* [2003] considered the prediction of the arrival time of interplanetary shocks and evaluated the performance of a shock prediction model with respect to a newly defined K_θ metric. As discussed in Appendix A, this metric is applicable for a different class of user than the one considered in this paper. *Bellanger et al.* [2003] considered the prediction of extreme changes in the ground magnetic field using a threshold algorithm in which the relative percentage of

[1]Laboratory for Atmospheric and Space Physics, University of Colorado, Boulder, Colorado, USA.
[2]Space Environment Center, National Centers for Environmental Prediction, National Weather Service, NOAA, Colorado, USA.
[3]High Altitude Observatory, National Center for Atmospheric Research, Boulder, Colorado, USA.

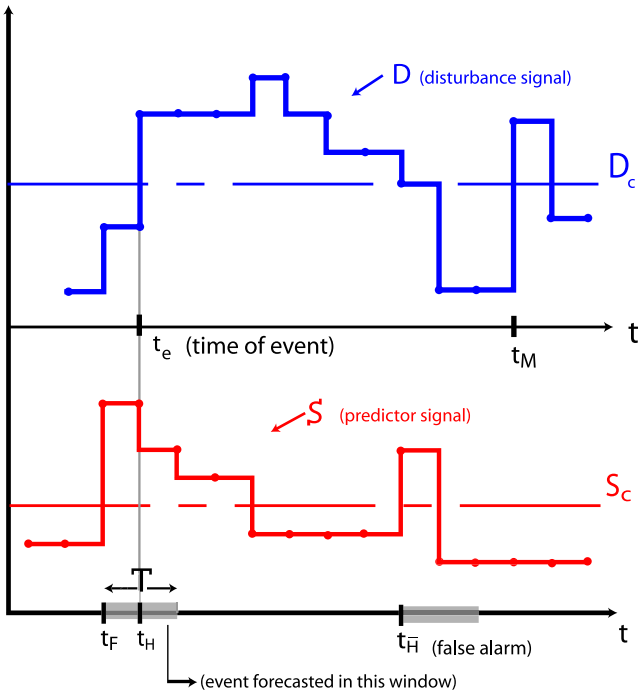

Figure 1. Schematic showing the various quantities used in this paper. The disturbance signal D can represent any quantity that is a proxy for system interruption or failure. The predictor signal S is derived from a model output. The time the forecast is made is t_F, and the time of a correctly predicted threshold crossing (or "event") of the disturbance signal D, t_e, is t_H. The time the forecast is extended over, T, is marked in gray. The two other times are the times of a false alarm, $t_{\overline{H}}$, and the time of a miss, t_M.

unpredicted events to the ratio of time covered by the alerts was plotted.

[5] *Gavrishchaka and Ganguli* [2001] developed a filter model that made use of a nonlinear transformation of solar wind measurements to predict the amplitude of the geomagnetic activity index, *AE*, only during times when it was above a critical level. The parameters of the model were determined by maximizing the data-model correlation for *AE* only on data above a critical value. The model was then evaluated for its ability to predict large events in terms of its false alarm and correct forecast predictions. This work clearly shows the problem faced by modelers. Namely, most model optimization techniques minimize a mean square error metric, even though the more user-appropriate metric involves the number of hits, misses, and false alarms.

[6] *Weigel et al.* [2003] considered an algorithm for predicting when the daily average flux of >2 MeV electrons would exceed a threshold value. The algorithm was evaluated according to the ratio of the number of correct forecasts to the number of false alarms. As discussed in the following section, a general class of users can determine if a model prediction has the potential to provide monetary utility with this ratio.

[7] From these recent works, it is difficult to tell what metric is most appropriate for event predictions. Here we attempt to synthesize and clarify the relationship between these analyses from the perspective of a hypothetical user of an event prediction. We begin with the framework provided by *Wilks* [2001] who developed the value score (VS) for a 2 × 2 contingency table. We extend this analysis and show that a more user-relevant presentation of binary event forecast results is not the VS, but rather the curves of the probability of a correct forecast, false alarm, and miss with respect to the amplitude of the disturbance being predicted at multiple threshold levels and realistic alert time spans. It is shown that these quantities can be used by a broad class of users to derive information that may be useful for analyzing or optimizing the economic utility of always taking action on the basis of a model forecast.

2. Definitions

[8] An event E is defined to be either a threshold crossing of a quantity or a situation where several quantities are in a certain range. This quantity does not have to be a direct measure of a variable that causes a system interruption or failure, but it must have some connection with interruption and failure, and should have a long time series available so that a statistical evaluation can be performed.

[9] Figure 1 shows a hypothetical disturbance and an event predictor time series. The time of the event is defined to be the first time interval in which the disturbance time series D crosses the critical level, D_c. These events are predicted by a model with an output of a scalar predictor signal, S, when it crosses the threshold value S_c, which can be adjusted along with the internal set of model parameters ($\equiv \mathbf{P}_m$).

[10] Some relevant quantities for determining the merit of a model that predicts an event are the statistical quantities defined in Table 1 and Figure 1. In Table 1 we have included definitions used in the previously mentioned works as well as the notation used in this article.

[11] For a binary forecast (event or no event predicted), all possible outcomes can be summarized in a 2 × 2 contingency table, as listed in Table 2. One problem with the 2 × 2 contingency table is that there are many ratios that can be formed, and different authors emphasize different ratios when presenting the prediction results of their models. However, as shown in the next section, only one ratio is important for a large class of users to determine if always following a forecast will have economic utility. In this paper we only consider predictions that can be summarized by a 2 × 2 contingency table, but this table can be generalized to cases where the quantity to be forecasted can take on many levels [*Doswell et al.*, 1990].

Table 1. Statistical Quantities Relevant for Event Forecasting

Symbol	Description
N_F	Number of forecasts
$N_{\overline{F}}$	Number of unit time intervals without a forecast
N_E	Number of events
$N_{\overline{E}}$	Number of nonevent intervals
N_H	Number of correct forecasts (“hits”)
$N_{\overline{H}}$	Number of forecast intervals during which no event occurred (“false alarms”)
R_F	Forecast ratio $\equiv N_H/N_{\overline{H}}$
N_M	Number of events not predicted (“misses”)
x	Number of nonevent and nonwarning intervals
N	Total number of events and nonevents(= $N_F + N_{\overline{F}}$, the total number of forecasts and nonforecasts)
C	Cost of a false alarm
B	Net benefit derived from a valid forecast
L_p	Loss that can be protected against, $B = L_p - C$
LR	Likelihood ratio
K	Loss structure ($K_{\min} = R_F^{-1}$)
Odds (z)	“Base rate” $\equiv P(z)/P(\bar{z})$)
$P(E)$	Probability of an event
$P(\overline{E})$	Probability of no event
$P(F\|E)$	Probability of a forecast given that there was an event
$P(F\|\overline{E})$	Probability of a forecasted event given that there was no event, that is, the false alarm probability

Alternatively, in section 3 we show how the need for contingency tables with dimensions greater than 2 × 2 can be eliminated for the class of user under consideration by using multiple threshold values.

[12] We have used definitions that allow an event forecast to be active for several time intervals. The total number of intervals N may not sum to the total number of data points; in this case care must also be taken in defining the cost, because for an extended warning interval the cost is with respect to taking mitigating action over the full warning interval.

3. Utility of a Forecast

[13] Our motivation is to introduce more user-relevant metrics than the standard correlation-based metrics with as little loss of generality as possible. We also seek to develop guidelines for those seeking to do space weather event prediction. We begin by making a few general assumptions about a user to restrict the problem. We assume the following.

[14] 1. The user takes the same mitigating action following each forecast of an event.

[15] 2. An “always mitigate” strategy yields a net monetary loss for the user.

[16] 3. The user seeks to maximize monetary gain, which we label as U_F, the utility of the forecast.

[17] The first assumption restricts the results to that produced by the forecast model output S and removes any influence of an intermediary decision maker. The second assumption restricts us to a set of systems for which continuous mitigation does not have utility. *Wilks* [2001] also considers the case where condition 2 is not satisfied; that is, always protecting the system yields a net benefit. This means that the optimal state of the system is always mitigated. In this case the utility should be considered as relative to this state as discussed in Appendix A.

[18] The two quantities most relevant for assessing the utility of forecast algorithm for users 1–3 are the number of correct forecasts (N_H) and the number of false alarm forecasts ($N_{\overline{H}}$), and the forecast algorithm has utility if U_F is greater than zero:

$$U_F \equiv BN_H - CN_{\overline{H}} > 0, \tag{1}$$

where C is the cost of taking mitigating action and B is the benefit from having taken mitigating action when an event occurred. The quantities N_M and x, which correspond to nonaction, do not enter into the equation because we are considering the utility with respect to a system that was never mitigated. The benefit can also be written as $B = L_p - C$, where L_p is the loss that is protected against. That is, the utility is with respect to a system that is never protected and thus sufferers a loss L_p for each event.

Table 2. Contingency Table Format Used in This Paper

	Observed		
Forecast	Yes	No	Total
Yes	N_H	$N_{\overline{H}}$	$N_F = N_H + N_{\overline{H}}$
No	N_M	x	$N_{\overline{F}} = N_M + x$
Total	N_E	$N_{\overline{E}}$	N

[19] Alternatively, the problem from can be framed from the prospective of losses. *Wilks* [2001] starts with L_F, the forecast loss

$$L_F = CN_F + L_pN_M = C\left(N_H + N_{\overline{H}}\right) + L_pN_M, \quad (2)$$

and demands that it is less than the loss expected from climatology $L_{\text{clim}} \equiv N_EL_p$. With $L_F < L_{\text{clim}}$ and the identity $N_H + N_M = N_E$, this equation is the same as equation (1).

[20] The user of an event forecast will want to choose the forecast model that maximizes U_F for their system. However, without knowledge of C and B, which are both user-dependent, the developer of a model can still determine for which users the model is potentially useful by reporting C/B ratios that satisfy

$$R_F \equiv \frac{N_H}{N_{\overline{H}}} > \frac{C}{B}. \quad (3)$$

If a user has a C/B ratio that satisfies this inequality, then taking action following every forecast will yield $U_F > 0$. If possible, further optimization of the model should seek to maximize U_F for those given values of C and B.

[21] The forecast ratio R_F is a useful metric for assessing if a forecast algorithm has the potential to provide economic utility for the user defined by conditions 1–3. However, we are considering a user who wants to maximize U_F. *Wilks* [2001] has introduced the value score (VS) metric for users 1–3. It is the ratio of the utility of the forecast to the utility of a perfect forecast

$$VS = \frac{1}{BN_E}\left(BN_H - CN_{\overline{H}}\right). \quad (4)$$

This value score is proportional to U_F and has the property of being unity for a perfect predictor ($N_H = N_E$, $N_{\overline{H}} = 0$) and has the advantage that it can be computed as a function of a single parameter, C/B.

4. Beyond the Value Score

[22] We have questioned if a correlation metric is sufficient for a user to make comparisons between predictors because a high data-model correlation does not necessarily imply maximum utility. The value score defined in the previous section is a metric that is of interest to users 1–3 because it is proportional to U_F. However, the value score may not be ideal for this user if the critical disturbance threshold D_c is a free parameter. Often the user-relevant optimization problem is to determine the values of the adjustable internal model parameters $\mathbf{P}_m$ and the D_c value that maximize U_F. (There is an additional free parameter, the time that the alert is extended over, T as illustrated in Figure 1. In the case of space weather forecasts, the range of values that T can take is usually restricted by the lead time provided by solar wind measurements or the time for a structure to propagate from the Sun to Earth. In the following, we assume that T is fixed.) In this section we consider the more realistic user that is described by conditions 1–4, where the additional condition is as follows.

[23] 4. The user has a C/B ratio that depends on D_c.

[24] If we are to compare two models at a given D_c, then the value score given by equation (4) is a suitable metric for the user defined by conditions 1–3. However, in a practical situation, B and C are likely to depend on D_c. For example, shutting down a power station to prevent damage from geomagnetically induced currents will result in a loss of income dependent on the extent of shutdown, while the amount of damage protected against depends on the magnitude of the geomagnetic event, represented by D_c.

[25] By rewriting equation (4) as

$$VS(\mathbf{P}_m, D_c) = \frac{U_F(\mathbf{P}_m, D_c)}{BN_E}, \quad (5)$$

it is clear that if B or N_E are dependent on D_c, then the values of $\mathbf{P}_m$ and D_c that maximize U_F will not necessarily be the same as those that maximize VS. (In general, B increases with D_c while N_E decreases with D_c, so the values of $\mathbf{P}_m$ and D_c that maximize U_F may be near the values that optimize VS in some cases. This possibility is considered in the following section.)

[26] The most general optimization problem for the user + modeler is to find parameters $\mathbf{P}_m$ and D_c that yield maximal utility

$$U_F(\mathbf{P}_m, D_c) = B(D_c)N_H(\mathbf{P}_m, D_c) - C(D_c)N_{\overline{H}}(\mathbf{P}_m, D_c). \quad (6)$$

We are left with the question of if a better metric of comparison and optimization can be devised for the users 1–4. Some of the choices are as follows.

[27] Option a: The modeler optimizes their model to maximize R_F.

[28] Option b: The modeler optimizes their model with respect to VS for several values of C/B.

[29] Option c: The users provide $C(D_c)$ and $B(D_c)$ curves. For each user the modeler determines an optimal $\mathbf{P}_m$ and D_c.

[30] Option d: The modeler reports many $N_H(\mathbf{P}_m, D_c)$ and $N_{\overline{H}}(\mathbf{P}_m, D_c)$ curves.

[31] Option e: The modeler reports $N_H(D_c)$ and $N_{\overline{H}}(D_c)$ curves for a fixed $\mathbf{P}_m$.

[32] How should modelers report results or optimize their model while still giving users 1–4 information that is of value? Options a and b are the most straightforward. Option a gives a user information they can use to determine if they can even benefit from always taking action following a forecast. Option b allows models to be more easily comparable side by side with a single number in the tradition of a correlation metric. A modeler can claim that "In this range of C/B values, my model is superior on the basis of the VS metric." However, for the realistic user

constrained by condition 4, the model may not be superior in a practical situation. Option c requires a user to provide information that may be of interest to their competitors. Option d puts a great burden on the modeler, especially if it takes a long time to compute a prediction time series.

[33] We suggest that the best compromise for users 1–4 is option e. With this information, a user can do a partial optimization by using the $N_H(D_c)$ and $N_{\overline{H}}(D_c)$ curves to maximize equation (6) with $\mathbf{P}_m$ constant. In the following section we show how a hypothetical user could use this information to determine the optimal parameter D_c.

[34] Adding the complication of allowing for a threshold-dependent B and C makes the analysis usable to a broader class of users than the value score. The drawback is that instead of determining a single set of model parameters $\mathbf{P}_m$ that optimize the value score for a fixed D_c, the modeler needs to provide either a curve that tells of the model performance as a function of D_c or needs to provide many models. Although we have added the complication of needing to present model performance as a function of D_c, we have simultaneously removed one of the restrictions inherent in using a 2×2 contingency table. By allowing D_c to vary, we are effectively considering a contingency table with more elements.

5. Examples

[35] In this section we give two examples of how a hypothetical user could use modeler-provided analysis to determine if the model's forecasts can benefit them. To simplify the presentation and analysis in this section, we restrict our analysis to that of a user that has, in addition to conditions 1–4, the following constraint.

[36] 5. The user has a cost C ($\equiv C_o$) that is independent of D_c.

[37] Given curves of the number of hits and false alarm forecasts, the user described by conditions 1–5 can compute

$$\frac{U_F}{C_o} = \frac{B(D_c)}{C_o} N_H(\mathbf{P}_m, D_c) - N_{\overline{H}}(\mathbf{P}_m, D_c) \tag{7}$$

on a grid of [B/C_o, D_c] values. To find an optimal value of D_c, the user plots their characteristic $B(D_c)$ curve on top of the U_F/C_o surface and locate the maximum U_F along the path.

[38] Because we are assuming that only results with fixed $\mathbf{P}_m$ are available, that is, the model is fixed, the user-relevant problem is to maximize

$$\frac{U_F}{C_o} = \frac{B(D_c)}{C_o} N_H(D_c) - N_{\overline{H}}(D_c). \tag{8}$$

Given the $N_H(D_c)$ and $N_{\overline{H}}(D_c)$ curves and the user's $B(D_c)/C_o$ versus D_c curve, U_F/C_o can plotted as a surface dependent on D_c and B/C_o. This surface can be compared to that of the value score

$$C_o VS = \frac{C_o}{B(D_c)} \frac{U_F}{N_E(D_c)}. \tag{9}$$

5.1. Prediction of MeV Electron Events

[39] In this section we consider the problem of forecasting when the dimensionless disturbance quantity $D \equiv J_e/(10^3$ particles sr^{-1} cm^{-2} s^{-1}) at $L = 4.4$ will cross a threshold value, J_{ec}, where J_e is the daily average fluence of energetic electrons measured by the PET instrument on SAMPEX. Although there is not a one-to-one correspondence between elevated J_e and satellite failure, J_e is a good proxy for failure or interruption in that long data sets of its measurements are available, and the probability of satellite failure or interruption is highly correlated with J_e [*Baker et al.*, 1987; *Vampola*, 1987].

[40] A simple prediction algorithm for J_e was analyzed by *Weigel et al.* [2003]. The algorithm states that if the daily average solar wind velocity, V, on day $t - 1$ was below V_c and above V_c on day t, then J is predicted to rise above a critical threshold level Je_c on day $t + 1$, $t + 2$, or $t + 3$. If the flux was pre-elevated ($Je(t) > Je_c$) then no prediction is made. In the analysis, the minimum usefulness ratio, R_F, was considered as a function of Je_c.

[41] The electron flux data are from the SAMPEX satellite while the daily averaged solar wind velocity data are from the OMNIWeb data set. The time interval of analysis starts on day 285 of the year 1994 and runs through day 365 of 2000, giving a total of 1951 days. (SAMPEX data are available from approximately 1993 on, but consistent near-Earth solar wind velocity data were not available in 1993 and most of 1994.)

[42] The number of hits, misses, and false alarm forecasts as a function of the critical disturbance level, Je_c is shown in Figure 2a for the algorithm with $V_c = 600$ km/s. The N_H and $N_{\overline{H}}$ curves in Figure 2a are used to compute the metrics R_F, VS, and U_F. For this example, we assume the user has a benefit/cost curve $B/C_o = 5 + 10 \tanh((Je/10^3 - 0.05)/0.05)$. The shape of this curve was chosen so that for large Je_c, B/C_o approached an asymptotic value, and for a small value of Je_c, B/C_o is zero. Such curves are user-dependent, and this form was chosen for illustration purposes only.

[43] The optimal values of Je_c determined for R_F, VS, and U_F derived in Figures 2a–2c differ substantially. The optimal value of R_F is determined by inspection from Figure 2a. The optimal values of VS and U_F are determined by locating their maximum values on the B/C_o curve shown in Figures 2b and 2c. The threshold value corresponding to maximal R_F, VS, and U_F are $Je_c/10^3 = 0.02$, 0.17, and 0.1, respectively. This result highlights the problem addressed in this article; If only R_F is presented, the user could not do the optimization in Figure 2c. If only the value score was presented, the user that chooses Je_c on

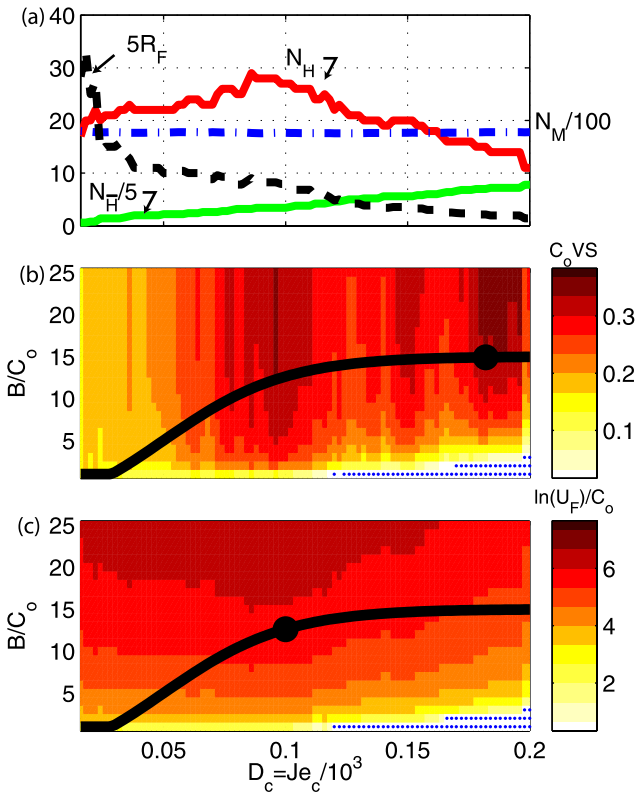

Figure 2. (a) Number of correct, false alarm, and missed forecasts and the ratio of correct to false alarm forecasts produced by the algorithm with V_c = 600 km/s that predicts the dimensionless disturbance $D \equiv Je/(10^3$ particles sr^{-1} cm^{-2}). J_e is the daily averaged energetic electron flux measured by the PET instrument on SAMPEX. (b and c) Value score, *VS*, and forecast utility, U_F, surfaces computed using equations (9) and (8), respectively, and the correct and false alarm curves in Figure 2a. The thick line in Figures 2b and 2c is the assumed B/C_o versus Je_c curve; the large dot is the maximum value of the surface along the line. In Figures 2b and 2c, small dots indicate negative values. (The R_F curve stops at 0.018 because we omit R_F values when the number of forecasts is less than 20 to prevent overfitting.)

that basis may obtain a result that differs from that if N_H and $N_{\overline{H}}$ curves were provided so that the user could compute U_F.

[44] Note that the surfaces shown in Figures 2b and 2c are jagged, which is most likely a result of the small number of events considered. For this reason any optimal value selected by the user would need to account for such uncertainty.

5.2. Prediction of Geomagnetic Storm Events

[45] In this section we consider predicting an excursion of the disturbance -*Dst* (on the basis of daily averaged *Dst*) above a threshold using only solar wind velocity measurements. The *Dst* data were obtained from OMNIWeb and the solar wind velocity data are the same as that used in the previous section.

[46] Statistically, the primary driver of *Dst* is the product of the solar wind velocity and the rectified north-south component of IMF (VB_s). (In that most of the variance in *Dst* can be explained by this product alone, even though many drivers exist that influence *Dst* that depend on other solar wind variables and combinations thereof [*Burton et al.*, 1975].) Long lead time prediction of B_z is much more difficult than prediction of V because B_z varies on a much shorter timescale. On 1-min timescales, *Chen et al.* [1997] used the fact that in a magnetic cloud, B_z is slowly varying, which allows its time evolution to be predicted when only a small fraction of the cloud has been observed. Here we suppose that only measurements of V are available and

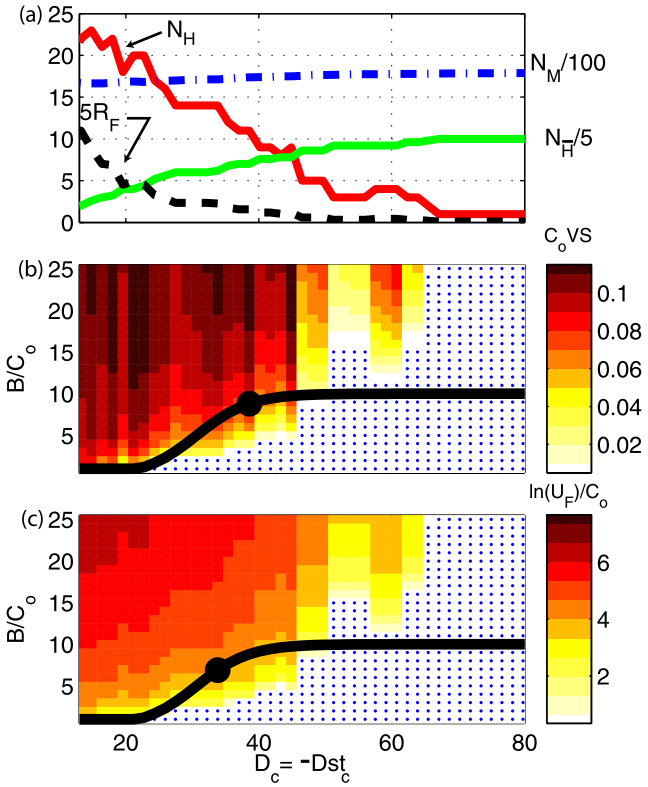

Figure 3. (a) Number of correct, false alarm, and missed forecasts and the ratio of correct to false alarm forecasts produced by the threshold algorithm with V_c = 567 km/s that predicts the disturbance -D_{st}. (b and c) Value score, *VS*, and forecast utility, U_F, surfaces computed using equations (9) and (8), respectively, and the correct and false alarm curves in Figure 3a. The thick line in Figures 3b and 3c is the assumed B/C_o versus Dst_c curve; the large dot is the maximum value of the surface along the line. In Figures 3b and 3c, small dots indicate negative values.

that the lead time of interest is one day. Note that by omitting B_z the number of misses and false alarms will increase significantly. However, for many purposes, depending on the user, even a seemingly poorly performing prediction algorithm may have some value given the alternative of no prediction at all.

[47] Figure 3a shows the curves of the number of correct and false alarm forecasts and the number of missed events of an algorithm that predicts an event in *Dst* using the same algorithm in the previous example. The N_H and $N_{\overline{H}}$ curves in Figure 3a are used to compute R_F, VS, and U_F, where we have assumed the user has benefit/cost curve $B/C_o = 5 + 5\ \tanh((D - 0.35)/0.05)$, where $D \equiv(-Dst - 25\ \text{nT})/158\ \text{nT}$. As in the previous example, the form of this curve was chosen to have a limiting value of B/C_o for large values of the threshold parameter, $-Dst_c$ and to be zero for a small values of the threshold parameter. Also, this curve was chosen so that the there was some maximal value of VS or U_F along its path.

[48] For the user path shown, the optimal Dst_c value is -38 nT, for the VS and -33 nT for U_F. The R_F metric again gives a much different answer; in this example the optimal value of *Dst* is -1 nT. As in the previous example, if the user is only given R_F as a function of a threshold parameter, their optimal value will be significantly different than what is determined if the full contingency table information is reported as a function of the threshold parameter.

[49] Note that the values of -33 nT and -38 nT represent very small geomagnetic storms and from Figure 3c the minimal value of Dst_c that yields positive U_F for any user is ~ -65 nT, which is a somewhat common occurrence (3.4% of the *Dst* values in the OMNIWeb data set from 1963–2002 fall below -65 nT). Given these numbers, it is quite likely that many users have a B/C_o curve that never yields a positive U_F, and hence are better off always ignoring the forecast. This is not surprising given the simplicity of the forecast algorithm that was used. Given the current state of solar wind velocity forecasting [*Baker et al.*, 2004], this is a reminder of the need for significant improvements in solar wind velocity predictions if extreme geomagnetic events are to be predicted at a level and lead time that is relevant to a potential user.

6. Summary and Conclusions

[50] We have presented some of the differences between what quantities a modeler evaluates, presents, and optimizes for model performance and what quantities are useful for a hypothetical user fitting the following description.

[51] 1. The user takes the same mitigating action following each forecast.

[52] 2. Both a coin flip forecast and an "always predict event" forecast yield a net monetary loss for the user.

[53] 3. The user seeks to maximize monetary gain.

[54] This user seeks to maximize the utility, U_F, which depends on the number of correct and false alarm forecasts.

[55] In the recent literature there have been numerous analyses of forecasts of rare events, all with emphasis on different forecast quality metrics. We have shown in two examples of how a modelers sometimes arbitrary choice of metrics to evaluate their model against may have a significant influence on a users decision. We have suggested that some of this influence can be eliminated if the researcher presents the curves $N_H(D_c)$, $N_{\overline{H}}(D_c)$, and $N_M(D_c)$, which are the number of a correct forecasts, the number of false alarm forecasts, and the number of misses as a function of the threshold quantity of the disturbance, D_c. With these curves, the optimal threshold value that a user selects is not influenced by the metric that a researcher decides to emphasize, comparison of model results may be more straightforward, and the metrics emphasized in the literature including R_F, K_θ, and many other ratios that can be derived from a 2×2 contingency table, can still be derived.

Appendix A: Appendix

[56] In this section we show how the formulations and metrics developed by *Wilks* [2001], *Mozer and Briggs* [2003], and *Matthews* [1997] are related to that developed in this article.

[57] *Wilks* [2001] considers the value score in the full range $0 \le C/L_p \le 1$. In the range of $0 \le C/L_p \le N_E/N$, the utility is with respect to the state of a system that is always mitigated, because if $N_E L_p \le NC$, then it follows that if the system is never mitigated there will be a loss of L_p for every event and that this loss is less than the loss incurred if the system is always mitigated. In this case the utility is with respect to the mitigated state and there is a gain for the unmitigated intervals of N-N_F and a loss of L_P for every miss

$$U_F = C(N - N_F) - L_p N_M, \tag{A1}$$

and the value score is

$$VS = \frac{(N - N_F)(C/L_p) - N_M}{N(C/L_p)(N - N_E)}. \tag{A2}$$

[58] In this paper we have only considered the utility for the case $N_E/N \le C/L_p \le 1$, because we required that the always predict and coin toss algorithms have $U_F < 0$. In this case the constraint is $C/L_p < N_E/(N_E + N_{\overline{E}})$ and the utility is with respect to the unmitigated state.

[59] *Mozer and Briggs* [2003] introduced and evaluated the metric

$$K_\theta = \frac{N_H(1-\theta) - N_{\overline{H}}\theta}{(N_H + N_M)(1-\theta)}, \tag{A3}$$

where

$$\theta \equiv \frac{C}{C + L_p}. \tag{A4}$$

This metric is equal to the value score when $C \ll B$, which can be seen by rewriting equation (A4) as

$$K_\theta = \frac{1}{BN_E}\left(BN_H - CN_{\overline{H}}(1 + C/B)^{-1}\right). \tag{A5}$$

Such a metric is probably more applicable to, for example, a patient who needs to decide if undergoing a medical testing procedure is useful since it deemphasizes the cost of a false alarm by a factor of $(1 + C/B)^{-1}$ relative to the value score. It is possible to derive this metric given the curves of $N_H(D_c)$, $N_{\overline{H}}(D_c)$, and $N_M(D_c)$, as proposed in this article.

[60] An additional formulation was given by *Matthews* [1997], who states that for a forecast to be useful,

$$\mathrm{LR} \cdot \mathrm{Odds}(E)K > 1, \tag{A6}$$

where $\mathrm{Odds}(E) = P(E)/P(\overline{E})$, and the loss ratio, LR, is defined as

$$\mathrm{LR} \equiv \frac{P(F|E)}{P(F|\overline{E})} = \frac{N_H}{N_{\overline{H}}}\frac{N_{\overline{H}} + x}{N_H + N_M} = \frac{N_H}{N_{\overline{H}}}\frac{N_{\overline{E}}}{N_E}, \tag{A7}$$

which follows from the relations in Table 1. With this, equation (A6) can be written as $R_F > K^{-1}$ which can be compared to equation (3), $R_F > C/B$, which represents the minimum hit to false alarm ratio that a model must have if it is to produce positive utility for a user with a give C/B ratio.

References

Baker, D., R. Belian, P. Higbie, R. Klebesadel, and J. Blake (1987), Deep dielectric charging effects due to high-energy electrons in Earth's outer magnetosphere, *J. Electrost.*, *20*, 3–19.

Baker, D. N., R. S. Weigel, E. J. Rigler, R. L. McPherron, D. Vassiliadis, C. N. Arge, G. L. Siscoe, and H. E. Spence (2004), Sun-to-magnetosphere modeling: CISM forecast model development using linked empirical methods, *J. Atmos. Sol. Terr. Phys.*, *66*, 1491–1497.

Bellanger, E., V. Kossobokov, and J. Le Mouel (2003), Predictability of geomagnetic series, *Ann. Geophys.*, *21*, 1101–1109.

Burton, R. K., R. L. McPherron, and C. T. Russell (1975), An empirical relationship between interplanetary conditions and *Dst*, *J. Geophys. Res.*, *80*, 4204–4214.

Chen, J., P. J. Cargill, and P. J. Palmadesso (1997), Predicting solar wind structures and their geoeffectiveness, *J. Geophys. Res.*, *102*, 14,701–14,720.

Doswell, I. C., R. Davies-Jones, and D. Keller (1990), On summary measures of skill in rare event forecasting based on contingency tables, *Weather Forecasting*, *5*, 576–585.

Gavrishchaka, V., and S. Ganguli (2001), Optimization of the neural-network geomagnetic model for forecasting large-amplitude substorm events, *J. Geophys. Res.*, *106*, 6247–6257.

Lindley, D. (1985), *Making Decisions*, 2nd ed., John Wiley, Hoboken, N. J.

Matthews, R. (1997), Decision-theoretic limits on earthquake prediction, *Geophys. J. Int.*, *131*, 526–529.

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

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

Vampola, A. (1987), Thick dielectric charging on high-altitude spacecraft, *J. Electrost.*, *20*, 21–30.

Weigel, R. S., A. J. Klimas, and D. Vassiliadis (2003), Precursor analysis and prediction of large-amplitude relativistic electron fluxes, *Space Weather*, *1*(3), 1014, doi:10.1029/2003SW000023.

Weigel, R. S., D. N. Baker, E. J. Rigler, and D. Vassiliadis (2004), Predictability of large geomagnetic disturbances based on solar wind conditions, *IEEE Trans. Plasma Sci.*, *32*, 1506–1510.

Wilks, D. (2001), A skill score based on economic value for probability forecasts, *Meterol. Appl.*, *8*, 209–219.

D. N. Baker and R. S. Weigel, Laboratory for Atmospheric and Space Physics, University of Colorado, Boulder, CO 80303, USA. (robert.weigel@lasp.colorado.edu)

T. Detman, Space Environment Center, National Centers for Environmental Prediction, National Weather Service, NOAA, W/NP9, 325 Broadway, Boulder, CO 80305, USA.

E. J. Rigler, High Altitude Observatory, National Center for Atmospheric Research, 3450 Mitchell Lane, Boulder, CO 80301, USA.