Evaluating Space Weather Forecasts of Geomagnetic Activity from a User Perspective

A. W. P. Thomson

British Geological Survey, West Mains Road, Edinburgh EH9 3LA, UK

Abstract. Decision Theory can be used as a tool for discussing the relative costs of complacency and false alarms with users of space weather forecasts. We describe a new metric for the value of space weather forecasts, derived from Decision Theory. In particular we give equations for the level of accuracy that a forecast must exceed in order to be useful to a specific customer. The technique is illustrated by simplified example forecasts for global geomagnetic activity and for geophysical exploration and power grid management in the British Isles.

1. Introduction

The concept of 'space weather' has come to be widely accepted as a description of the current state of the space environment, 'as a result of dynamical solar, heliospheric and magnetospheric phenomena in the context of potential impacts on technological systems'. (Quotation taken from the executive summary of [*ESA*, 1999].) Space weather therefore implies active consideration of the impact of the space environment on human technologies, for example, satellites, power grids, communications and geophysical exploration.

There exist numerous methods of predicting changes in the near-Earth environment (as can be seen in, for example, [*ESA*, 1999]). However it is essential to be able to 'translate' these forecast data, for example radiation belt populations or geomagnetic indices, which may well have an obvious meaning to the forecaster, into appropriate formats that can be digested by the public, industry and commerce. We also need techniques that can properly evaluate the relative costs and benefits of using forecasts, again in appropriate formats. In this paper we attempt to address some of these issues.

Forecasts of geomagnetic activity, for example by the Space Environment Center (SEC) and the Regional Warning Centers, are typically based on an expert assessment of various space weather data. These data may include reports of solar and geophysical activity and satellite observations of the Sun and of interplanetary space. Linear and non-linear models and forecasts of magnetic and other data may also be used. The forecaster aims to go beyond the accuracy obtained by numerical schemes by adding experience and inference.

In this paper we examine a simple 'test forecast' of geomagnetic activity, based on the daily coronal disturbance and solar active region reports from SEC (the terms used here are as given in [*Heckman*, 1988] and are widely used in the Space Weather community). If solar type II or type IV radio emission is reported, on a given day, along with at least *C*-class X-ray activity and reports of 10 cm and 245 MHz

emission, we say that a 'forecast' of geomagnetic activity has been made, for two to four days ahead. This 'forecast' is taken to mean the expectation of geomagnetic activity above certain thresholds that are defined below. The motivation for using these particular data as surrogates for coronal mass ejection is suggested by, for example, [*Luhmann*, 1997; *Jackson*, 1997] and references therein. A combination of precursors is used merely to increase the confidence with which one can say that a potentially geo-effective mass ejection has occurred. This gives a rather 'mechanical' forecast that one would expect to be out-performed by a human forecaster, applying experience and inference to a wider data set. But it is a simple example for the purposes of this study and is easily retrospectively analyzed for discussion of such issues as accuracy and false alarm rate. It is, however, worth emphasizing that this is an *example* of a forecast and is not one in use by SEC or any other institute. Further, an expert forecaster would ideally expect to produce fewer false alarms; a key aspect of the results we present here.

To quantify the value of this simple forecast system, we apply Decision Theory [e.g. *Matthews*, 1996, 1997]. Decision Theory provides a contrasting perspective on forecast usefulness in comparison with existing metrics, for example those described by [*Detman and Vassiliadis*, 1997] and [*Detman and Joselyn*, 1998].

We investigate the method with reference to global geomagnetic activity, represented by the magnetic index Ap, and to local activity that might affect oil and gas drilling operations in the North Sea and power grid management in the UK. In each case an example 'action threshold' of geomagnetic activity has been determined, partly by previous study and experience of geomagnetic activity, both globally and in the UK, and partly in relation to the impact on technologies, where known [e.g. *Beamish* et al., 2000; *Kappenman*, 1996]. Discussions with industry representatives have also been helpful. Thus, the particular thresholds used here are in effect proxy data for the technological impact of space weather and may not be in a one-to-one correspondence with, for example, dates and times of specific problems experienced by particular industries. Moreover the activity levels used are 'guide' values and may change over time, as the science and our understanding of space weather develops. However the methodology presented does provide an example of how one might more closely link the science and data of space weather to the day-to-day decisions that face potential forecast users.

2. Decision Theory and Applications in Geomagnetic Storm Forecasting

In the following we define A as the occurrence of geomagnetic activity above some threshold (at various levels, depending on the end-user, and defined below), F is a forecast

of geomagnetic activity based on solar observation, in the sense described earlier, and D is the action or decision of the user in response to the forecast. A user action might include protecting sensitive equipment or postponing some activity, both of which imply a cost. Pr denotes a probability, with a vertical bar expressing the term 'given' in a joint probability; '~' denotes the negation of any of A, F or D; and '&' denotes the Boolean 'and'. Depending on A or ~A, we can construct various probability combinations. For example, $Pr(\sim A \mid F)$ would denote the probability that no activity above the threshold was recorded given that a forecast of geomagnetic activity (above threshold) had been made for the period. $Pr(\sim A \& D)$ denotes, for example, the probability that the user takes some (positive) action and that no activity occurs.

A pure number, L_{tot}, gives the total loss (i.e. cost or utility to the user) associated with the specific combinations of geomagnetic activity and user actions:

$$L_{tot} = L_{00}Pr(A \& D) + L_{11}Pr(\sim A \& \sim D)$$
$$+ L_{10}Pr(\sim A \& D) + L_{01}Pr(A \& \sim D) \qquad (1)$$

The L_{ij} $(i,j = 0,1)$ can be specified by the user and may represent a simple ranking of the possible outcomes or, perhaps more significantly, a monetary value associated with each action. The main costs to the user are represented by the 'off-diagonal' terms L_{10} and L_{01}.

The various joint probabilities in L_{tot} can be expanded on F and ~F to give terms that reflect the accuracy of the alert and the users' reaction to it: see [Matthews, 1997] for details. For example, the likelihood ratio, LR, is a common measure of the inherent accuracy of the alert:

$$LR = Pr(F \mid A)/Pr(F \mid \sim A) \qquad (2)$$

$LR>1$ implies that the forecast method has merit. LR answers the question as to whether a forecast was made prior

to each event of 'activity' or 'no activity'. Also, $Pr(D \mid F)$, $Pr(D \mid \sim F)$ capture the user's response to forecasts. Ideally, the user has complete confidence in the forecast and, as such, $Pr(D \mid F) = 1$, $Pr(D \mid \sim F) = 0$.

An important quantity - determined by the user - is the loss structure, K, based on the L_{ij}:

$$K = (L_{01} - L_{00})/(L_{10} - L_{11}) \qquad (3)$$

This measures the relative loss due to complacency (i.e. doing nothing before significant activity occurs) and false alarms (doing something in response to a forecast of activity though no significant activity subsequently occurs).

The criterion for judging an alert, or forecast, F, is that the total loss of taking F seriously is less than the total loss that arises from ignoring it. It can then be shown using Bayes Theorem [Matthews, 1997]), from $L_{tot}/(L_{10} - L_{11})$, and keeping terms that summarize the user response to the forecast (i.e. $Pr(D \mid F)$ and $Pr(D \mid \sim F)$), that a user is justified in making a decision based on receipt of F (i.e. $Pr(D \mid F) = 1$, $Pr(D \mid \sim F) = 0$) when

$$LR \times Odds(A) > 1/K \qquad (4)$$

Where $Odds(A) = Pr(A)/Pr(\sim A)$, is the base-rate odds for geomagnetic activity above the given threshold.

Equation (4) is a new criterion for assessing the value of forecasts. A user with cost structure K will save money by acting on forecasts if (4) is satisfied for those forecasts. If A is an existing activity measure (e.g. a well known magnetic index) the left side of (4) is readily calculated from an historical database of A and a record of forecasts of A. If A is a new measure of activity, perhaps specifically tailored for a new user, then $Odds(A)$ may still be readily determined (e.g. when A is a function of existing geophysical data). LR, however, depends on the forecasting of events whose
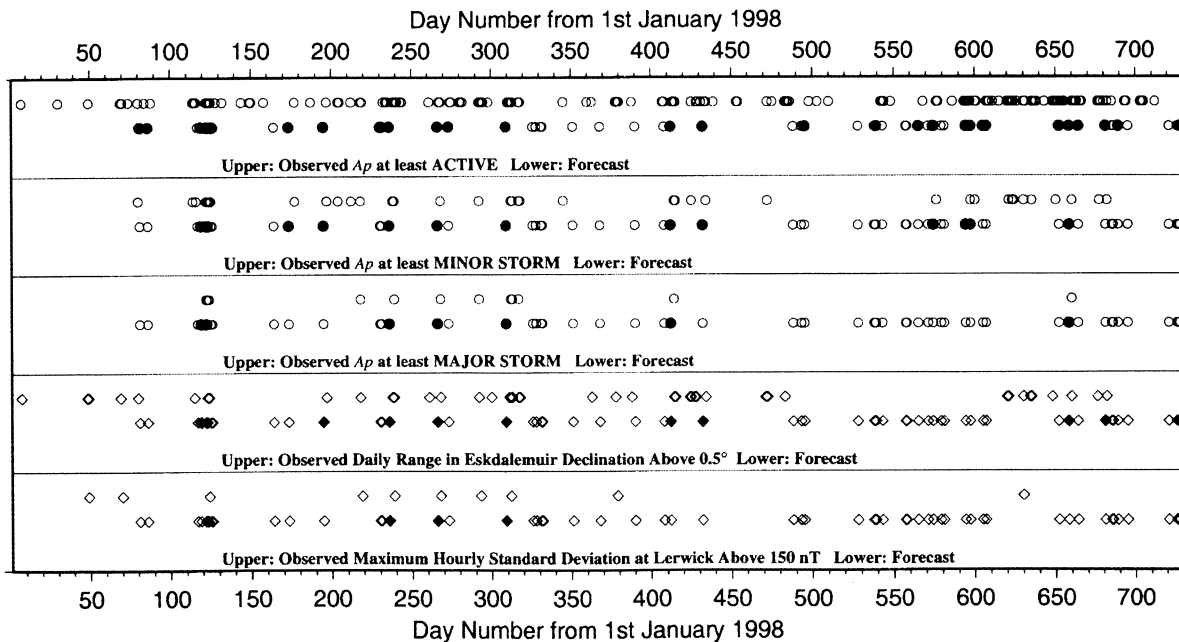

Figure 1. Days when geomagnetic activity exceeded the thresholds described in the text are shown by symbols in the upper time series in each of the five boxes. The lower time series in each box are days on which forecasts were made, using the simple solar activity test described in the text. Shading highlights forecasts that were followed by geomagnetic activity within the time frame of the forecast.

Table 1. Likelihood Ratios and Loss Structures for the Data of Figure 1.

Effect Measured = Q	Number of days with $Q \geq$ threshold		Number of days with $Q <$ threshold		$Pr(F\|Q{\geq}T)$	$Pr(Q{\geq}T\|F)$	$Pr(Q{\geq}T)$	LR	K
	# days with alert	# days with no alert	# days with alert	# days with no alert					
Global Activity									
= Active	47	102	100	484	0.315	0.320	0.204	1.84	2.12
= Minor Storm	18	20	129	566	0.474	0.122	0.052	2.55	7.15
= Major Storm	9	3	138	583	0.750	0.061	0.016	3.92	15.69
UK Activity									
Decl. at ESK $\geq 0.5°$ (drilling)	15	29	132	557	0.341	0.102	0.060	1.77	8.85
HSD at LER ≥ 150 nT (power-grid)	4	6	143	581	0.400	0.027	0.014	2.03	34.69

Various forecast and event data associated with the likelihood ratios, *LR*, and loss structures, *K*, for the data shown in Figure 1. Note that *K* is the minimum value that a user must have for the forecast to be useful. There were 733 days in the sample, 147 days of which were covered by 57 separate (occasionally overlapping) forecasts. 'Decl' means the daily range in geomagnetic declination and HSD is the hourly standard deviation in the horizontal field. Q represents a geomagnetic quantity, used here as a surrogate for industry 'problem events'. *T* is a threshold for Q. Activity '*A*', as used in the text, is here equivalent to '$Q{\geq}T$'.

outcomes must not be known, and may not be easily calculable retrospectively without bias, if human interpretation is part of the process. Where human intervention is not part of the forecast process (e.g. an automated forecast, such as here), then this concern probably does not arise. The left side of equation (4) may be further simplified, through Bayes rule, to give

$$Odds(A \mid F) > 1/K \qquad (5)$$

Which establishes $Odds(A|F)$ as a practical metric of value for forecasts, *F*.

In Figure 1 we show time series of dates on which global geomagnetic activity exceeded the three thresholds of active conditions, and minor and major storm levels (respectively $Ap{>}15$, $Ap{=}30$ and $Ap{=}50$) in 1998 and 1999. We also indicate those dates when the daily range in declination at Eskdalemuir magnetic observatory exceed $0.5°$ and when the hourly standard deviation (HSD) in the horizontal component of the field exceeded 150 nT at Lerwick observatory. These geomagnetic thresholds are potentially significant for, respectively, directional drilling concerns in the North Sea (from discussions with industry representatives) and for induced currents in the UK power grid [from *Beamish* et al., 2000]. Against each of these five time series we indicate dates on which geomagnetic activity forecasts were made, according to the solar activity criteria defined earlier. Note that the same forecast series is used for each of the five activity types. The forecast is taken to mean geomagnetic activity will occur, above threshold, on one or more days, between two and four days after the forecast date. Shaded forecast symbols indicate those forecasts that were therefore 'successful' in this context. It can be seen that there is some degree of correlation between dates of forecasts and some events in the geomagnetic activity time series. For active Ap (top box of Figure 1) we have many events but relatively few forecasts, whilst at high HSD (lower box of Figure 1) there are few threshold crossings but relatively many forecasts.

Decision Theory is now used to properly quantify the success level and hence supply a measure of the value of these example forecasts for each user group.

In Table 1 (see caption) significant data and conditional probabilities are described, as well as *LR* and a 'theoretical' *K*, calculated from equation (4) assuming equality, for tests of this forecast series against the five thresholds. Note that the first four data columns of Table 1 comprise a contingency table for each row in the sense of, for example, [*Detman and Joselyn*, 1998]. In all cases $K > 1$, which implies that, for this forecast 'service' to be useful to a user, the cost of complacency must always exceed that of the cost of reacting to false alarms. *K* roughly increases in inverse proportion to the *a priori* likelihood of the event. Typical loss structures in weather forecasting are of the order of 3-10 [*Matthews*, 1997]. Thus this geomagnetic forecast 'service' provides a comparable cost/benefit level except perhaps at the lowest levels of geomagnetic activity.

In general, any user has to decide on an appropriate response to geomagnetic forecasts which might warn of severe activity and which have, either way, cost implications. As an example, a false alarm (i.e. reacting to an erroneous alert) might cost a power company $X every time, in terms of manpower and equipment protection, etc. Ignoring a correct forecast then needs to have the potential (in theory) to cost greater than about $35X, in comparison (final entry of Table 1). Whether the company can do anything, that is make use of the forecast service as it stands, in the event that its' *true* loss structure is less than $35X, e.g. partial protection of some part of its operation, cannot be decided here.

We can see that, as the event being forecast becomes *a priori* rarer, e.g. at higher Ap thresholds, *LR* increases, that is, the method is more accurate. This may have been anticipated: the largest geomagnetic storms are generally preceded by quite marked solar precursors. Only at the highest activity thresholds does this 'service' compare reasonably with a UK weather forecast *LR* of about 5 [*Matthews*, 1997]. Further increases in *LR* are most likely to follow from a reduction in

the false alarm rate. A more sophisticated forecast service could be tailored for this purpose. Finally, [Matthews, 1996] discusses an important general issue: the significance of low base-rates for events and the effect of this on the public perception of forecast accuracy.

3. Discussion and Conclusions

Decision Theory enables potential users of space weather forecasts to assess the relative cost/benefit of acting upon a forecast of an event that might affect their particular interests. It is therefore a useful addition to existing techniques [e.g. Detman and Vassiliadis, 1997; Detman and Joselyn, 1998].

A basic requirement of the user is that they maintain some long-term time series (dates and times) of 'problem' events, for example, transformer damage in the power industry, or satellite anomalies, which can then be related to, for example, the geomagnetic record.

For the forecaster, in order to improve on current forecasts of space weather, particularly forecasts from numerical prediction schemes, well-correlated solar, interplanetary and geomagnetic data are desirable. However some caution should be exercised in accepting at face value the apparent significance level of prospective precursor data. The method of retrospective validation of time association between precursors and activity may prove useful here [Mulargia, 1997; Cox, 1955]. This method can take into account the biases inherent in simple statistical correlations between data sets and can provide a clearer statement of the true value of individual precursors. It does not appear that retrospective validation has been widely used in the field of solar and geomagnetic activity, although it has been applied in geophysics [Mulargia, 1997].

For example, we currently use a solar precursor time series, a daily solar index SI [Thomson, 1996]. SI is a weighted sum of those SEC reported solar data (e.g. type II and IV radio bursts, radio emissions at various wavelengths, numbers of energetic flares and active regions), which can show a reasonable numerical correlation with the Ap index. The purpose of SI is to add information to a neural net predictor of Ap, particularly since the Ap auto-correlation, the basis of many linear and non-linear prediction schemes for this index, is significant only for about 1-3 days forward lag. SI has shown a cross-correlation with Ap of up to 0.6 at 3-4 days ahead in some years since 1992, though this has been seen to reduce to less than the level of the Ap auto-correlation around solar minimum. Where there is a significant cross-correlation, SI would appear to be of value in improving forecasts of Ap, particularly at the longer lags. Retrospective validation has been applied to the visual associations seen between the peaks in both time series. In a preliminary study we have found that any such 'precursor-then-event' correlation is significant only between the most active SI and Ap, i.e. between

approximately the top 5% of activity in each time series, and only then during the most active years of the solar cycle. It thus seems likely that high numerical cross-correlations are dominated by the magnitude of the relatively rare, high activity events. Even so, the Ap neural net predictor with SI input does perform better than Ap self-prediction neural nets.

References

Beamish, D., T. D. G. Clark, E. Clarke, A. W. P. Thomson, Geomagnetically induced currents in the UK: geomagnetic variations and surface electric fields. *J. Atmos. Sol. Terr. Phys.* to appear, 2000.

Cox, D. R., Some statistical methods connected with series of events, *J. R. Stat. Soc.*, B, *17*, 129-164, 1955.

Detman, T. and J. A. Joselyn, Real-time Kp predictions from ACE real time solar wind, *SOLAR WIND NINE, Proceedings of the Ninth International Solar Wind Conference, Nantucket, MA*, October 1998, AIP Conference Proceedings 471, eds. Shadia Rifai Habbal, Ruth Esser, Joseph V. Hollweg, Philip A. Isenberg, 729-732, 1998.

Detman, T. D. and D. Vassiliadis, Review of techniques for magnetic storm forecasting, in *Magnetic Storms*, eds. B. T. Tsurutani, W. D. Gonzalez, Y. Kamide and J. K. Arballo, AGU Monograph, *98*, 253-266, 1997.

ESA, Workshop on space weather, *WPP-155*, ESTEC Conference Bureau, PO Box 299, 2200 AG, Noordwijk, Netherlands, 528 pp, 1999.

Heckman, G.R., Space Environment Center Glossary of Solar-Terrestrial Terms (also at http://www.sec.noaa.gov/info /glossary.html). *NOAA Special Publication*, 69 pp, 1988.

Jackson, B. V., Heliospheric observations of solar disturbances and their potential role in the origin of geomagnetic storms, in *Magnetic Storms*, eds. B. T. Tsurutani, W. D. Gonzalez, Y. Kamide and J. K. Arballo, AGU Monograph, *98*, 59-76, 1997.

Kappenman, J. G., Geomagnetic storms and their impact on power systems, *IEEE Power Engineering Review*, 5-8, May, 1996.

Luhmann, J. G., CMEs and space weather, in *Coronal Mass Ejections*, eds. N. Crooker, J. A. Joselyn and J. Feynman, AGU Monograph, *99*, 291-299, 1997.

Matthews, R. A. J., Why is weather forecasting under a cloud?, *Mathematics Today, 32*, 168-170, 1996.

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

Mulargia, F., Retrospective validation of the time association of precursors, *Geophys. J. Int., 131*, 500-504, 1997.

Thomson, A. W. P., Non-linear predictions of Ap by activity class and numerical value, *Pure Appl. Geophys., 146*, 163-193, 1996.

A. W. P. Thomson, British Geological Survey, West Mains Road, Edinburgh EH9 3LA, Scotland, UK. (e-mail:a.thomson@bgs.ac.uk)