Solar Physics
DOI: 10.1007/•••••-•••-•••-••••-•

Nowcasting Solar Energetic Particle Events Using Principal Components Analysis

A. Papaioannou[1], A. Anastasiadis[1], A. Kouloumvakos[2], M. Paassilta[3], R. Vainio[3], E. Valtonen[3], A. Belov[4], E. Eroshenko[4] M. Abunina[4] A. Abunin[4]

Abstract In this work, we perform a principal component analysis (PCA) on a set of six solar variables (*i.e.* CME width/size (s) and velocity (u), logarithm of the solar flare (SF) magnitude (log$SXRs$), SF longitude (lon), duration (DT) and rise time (RT)). We show a classification of the solar energetic particle (SEP) events radiation impact (in terms of the NOAA scales) with respect to the characteristics of their parent solar events. We further attempt to infer the possible prediction of SEP events. In our analysis, we utilize 126 SEP events with complete solar information, from 1997 to 2013. Each SEP event is a vector in six dimensions (corresponding to the six solar variables used in this work). PCA transforms the input vectors into a set of orthogonal components. The mapping of the characteristics of the parent solar events to a new base defined by these components led to the classification of the SEP events. We further, applied logistic regression analysis with single, as well as, multiple explanatory variables, in order to develop a new index (I) for the nowcasting (short-term forecasting) of SEP events. We tested several different schemes for I and validated our findings with the implementation of categorical scores (Probability of Detection - POD, False Alarm Rate - FAR). We present and interpret the obtained scores and we discuss the strengths and weaknesses of the different implementations. It was

[1] Institute for Astronomy, Astrophysics, Space Applications and Remote Sensing (IAASARS), National Observatory of Athens, I. Metaxa and Vas. Pavlou St., GR-15236, Penteli, Greece.
[2] IRAP, Universit de Toulouse, CNRS, CNES, UPS, Toulouse, France
[3] Department of Physics and Astronomy, University of Turku, 20014 Finland.
[4] Institute of Terrestrial Magnetism, Ionosphere and Radiowave Propagation by N.V. Pushkov RAS (IZMIRAN), Moscow Troitsk, Russia.

shown that I holds prognosis potential for SEP events, with the maximum POD achieved being 77.78% and the relative FAR being 40.96%.

Keywords: Solar Energetic Particle Events; Statistical Methods; Flares; Coronal Mass Ejections; Principal Components Analysis, Logistic Regression Method

1. Introduction

Solar energetic particle (SEP) events are marked as sudden excesses over a background level in the time profiles of several different energies, ranging from $\approx$ 10 keV to $\approx$ 10 GeV/nuclei. Those last from hours to a few days and include electrons, protons, alpha particles, and heavier ions up to Fe (Reames, 2017). SEP events are categorized into "impulsive" and "gradual" ones based on their parent solar events (Reames, 1999). In particular, the impulsive SEP events are considered to be associated with solar flares (SFs) and type III radio bursts. These events have an Fe/O $\approx$1 and a narrow injection cone. On the other hand, gradual SEP events are presumably associated with coronal mass ejections (CMEs) and type II radio bursts, while they have an Fe/O ratio $\approx$0.1 and a wide injection cone (Reames, 2013). The underlying argumentation for this dichotomy is based on the fact that SEP events are produced either in the solar atmosphere by particle acceleration processes in association with flares of class higher than C (Anastasiadis, 2002) or by a CME-driven shock in the interplanetary (IP) space (Cane and Lario, 2006). However, as the observational evidence at hand shows, this dichotomy has been regularly violated (Kocharov and Torsti, 2002; Cane, Richardson, and Von Rosenvinge, 2010; Papaioannou *et al.*, 2016). At this point, it is worth noting that recent identifications of wide-spread SEP events (*e.g.* Rouillard *et al.*, 2012; Dresing *et al.*, 2012; Kouloumvakos *et al.*, 2016) has challenged and extended our current understanding on SEP events (Dröge *et al.*, 2010; Wiedenbeck *et al.*, 2012; Gómez-Herrero *et al.*, 2015; Lario *et al.*, 2016, 2017).

SEP events cause failures to spacecraft, damaging their electronic components (Iucci *et al.*, 2005; Mikaelian, 2009) and, at the same time, pose a radiation threat for astronauts (Turner, 2006; Chancellor, Scott, and Sutton, 2014) and air crews (Lim, 2002; Mishev, 2014; Tobiska *et al.*, 2015). As a result different concepts and techniques, focused on the short-term forecasting (nowcasting) of SEP events, have been developed and set to operation by the scientific community. As a rule, these concepts are based on data driven approaches. The basic inputs utilized are: the magnitude and the position of the parent SF on the solar disk (Smart and Shea, 1989), the time-integrated soft X-ray flux of SFs, and the occurrence (or non-occurrence) of metric radio type II and type IV bursts (Balch, 1999, 2008), evidence of particle escape (*i.e.* type III bursts) (Laurenza *et al.*, 2009; Alberti *et al.*, 2017), near-Earth differential and integral proton fluxes (Núñez, 2011), type II and type III radio bursts (Winter and Ledbetter, 2015). In addition, the scatter-free propagation of the near-relativistic electron measurements or of the sub-relativistic protons (E$\geq$433 MeV) have been utilized either to infer the corresponding intensity of ions in the IP space (Posner, 2007) or to develop a concept for the prompt identification of ongoing high energy SEP events (Souvatzoglou *et al.*, 2014). Nowadays, the need for integrated SEP event nowcasting systems has led to the implementation of ensemble solutions, among them: the Forecasting Solar Particle Events and Flares (FORSPEF) tool (Papaioannou *et al.*, 2015; Anastasiadis *et al.*, 2017) and the Space Radiation Intelligence System (SPRINTS) framework (Engell *et al.*, 2017).

A wealth of statistical studies has indicated the dependence of the probability of occurrence of SEP events on the magnitude and the longitude of the SF (Kurt *et al.*, 2004; Belov *et al.*, 2005; Belov, 2009), and the relation between the peak proton flux and the velocity of the CME (Kahler, 2001), as well as the magnitude of the SF (Cane, Richardson, and Von Rosenvinge, 2010). It has also been shown that SEP events are related to both type II and type III radio bursts (Miteva, Samwel, and Krupar, 2017). However, most studies are limited to two dimensional (2D) correlations. In addition, similar coefficients are identified for the pair-wise correlation of the SEP peak intensity (at E>10 MeV) to both the SF magnitude and the CME speed (Dierckxsens *et al.*, 2015; Papaioannou *et al.*, 2016; Paassilta *et al.*, 2017; Belov, 2017), while the situation is further complicated by the fact that the solar parameters are not independent. To this end, Trottet *et al.* (2014) performed an analysis with partial correlation coefficients in order to disentangle the effects of correlations between the solar parameters themselves. The next step was to investigate possible 3D relationships among three numeric variables projected in two dimensions. From such a study it was verified that the combination of strong SFs and fast CMEs result in enhanced radiation storms. Furthermore it was shown that strong SFs result in enhanced radiation effects even when associated with moderate CMEs. In addition, these strong SFs can lead to major radiation storms even when they are not situated on the west part of the visible solar disk (Papaioannou *et al.*, 2016). Therefore, aiming at a higher dimensional order correlations seems to be the way forward. Given the complexity of the parent solar events of SEPs (*e.g.* SFs, CMEs) and the different variables (*e.g.* Geostationary Operational Environmental Satellite (GOES) peak photon flux, longitude of the SF, velocity and width of the CME) that give rise to their peak proton flux, possible new methods for the nowcasting of SEP events have to be associated with more accurate mathematical methods of statistical analysis.

To this end, one method that can be used is the principal component analysis (PCA), a multivariate statistical technique being used to examine the interrelations among a set of variables (*e.g.* a *dataset*) aiming to identify the underlying structure of those variables (Jolliffe, 2002). In particular, it extracts the essential information hidden in the dataset, represents it as a set of new orthogonal variables - called *principal components* (PCs), and displays the pattern of similarity of the observations and of the variables as points in maps (Abdi and Williams, 2010). PCA has often been used in several diverse scientific fields, since it is a straightforward, non parametric method of extracting relevant information from multi-variable data sets (Shlens, 2014). Recently, this method was applied to radio data (*i.e.* type II and type III bursts identifications) and was proven to lead to promising results (Winter and Ledbetter, 2015). Thereby, the goal of this article is to utilize PCA in order to perform a classification and to derive and test a possible index (I) for the nowcasting of the SEP events.

2. Data and Methods

2.1. A Database of SFs, CMEs and SEP Events

Recently, we presented a new catalogue of SFs, CMEs, and SEP events, spanning over almost three solar cycles from 1984 to 2013 (Papaioannou *et al.*, 2016). This database includes a total of 20498 SFs, 3680 CMEs, and 314 SEP events[1]. The relevant solar information incorporated in the catalogue (for both SEP and non-SEP events), comprises: a) peak soft X-ray (SXR) flux, b) longitude, c) latitude, d) SXR fluence, e) rise time, and f) duration of the parent SF, as well as, g) the velocity and h) the width of the associated CME. For the SEP events, the peak proton flux and the fluence were determined for four integral energy channels (E>10-, >30-, >60- and >100 MeV) for all SEP events with a peak proton flux, at E > 10 MeV, of > 1 pfu (pfu = particle flux unit = particle cm^{-2} sr^{-1} s^{-1}). In order to apply principal component analysis (PCA), we have identified a complete parametric grid of six (6) solar variables (*i.e.* CME width/size (s) and velocity (u), logarithm of the SF magnitude ($\log SXRs$), SF longitude (lon), duration (DT)[2] and rise time (RT)), from the aforementioned database, covering the time period from 1997–2013. This resulted in a total of 3663 records with complete information for all six variables, out of which 126 were SEP events and 3537 were non-SEP events.

2.2. Principal Component Analysis (PCA)

The principal component analysis (PCA) is a multivariate technique that allows the analysis of a data table in which observations are described by several inter-correlated quantitative dependent variables (Abdi and Williams, 2010). The goal of traditional PCA is to: (a) reduce the number of variables and (b) detect structures in the relationships between variables, that is to classify variables. As concerns (a), PCA reduces the number of variables to a smaller number of uncorrelated variables called *principal components* which account for, as much as possible, variance in the data. By definition, the first principal component (PC1) is the one which maximizes the variance when data are projected onto a line and the second one (PC2) is orthogonal to PC1 while it still maximizes the remaining variance.

Mathematically speaking, PCA is defined as an orthogonal linear transformation that transforms the set of initial variables to a new coordinate system such that the greatest variance by some projection of the data lies on the first coordinate which is called the first principal component (PC1), the second greatest

[1] The associated CMEs span from 1997 to 2013, with the availability of the continuous SOHO/LASCO measurements

[2] The start time of an X-ray event is defined as the first minute, in a sequence of 4 minutes, of steep monotonic increase in the 0.1-0.8 nm flux. The end time is the time when the flux level decays to a point halfway between the maximum flux and the pre–flare background level. Thereby, the duration time (DT) is the time difference between the end and the start time of the flare.

variance on the second principal component (PC2), and so on (*e.g.* PC3, PC4, ...) (Shlens, 2014).

In the most general case, a PCA transformation is defined by a set of p-dimensional vectors (p is the number of variables under study) of loadings $\mathbf{w_{(k)}}$ (k is the number of the component) that map each row vector of the initial variables $\mathbf{X_{(i)}}$ to a new vector of principal component scores $\mathbf{t_{k(i)}}$, given by:

$$\mathbf{t_{k(i)}} = \mathbf{X_{(i)}} \cdot \mathbf{w_{(k)}}, \tag{1}$$

in such a way that the individual component scores $\mathbf{t}$ inherit the maximum possible variance from $\mathbf{X}$, with each loading vector $\mathbf{w}$ constrained to be a unit vector. In order to maximize the variance, the loading vectors $\mathbf{w_{(k)}}$ have to satisfy the following criterion:

$$\mathbf{w}_{(k)} = \underset{\|\mathbf{w}\|=1}{\arg\max} \left\{ \|\hat{\mathbf{X}}_k \mathbf{w}\|^2 \right\} = \arg\ \max \left\{ \frac{\mathbf{w}^T \hat{\mathbf{X}}_k^T \hat{\mathbf{X}}_k \mathbf{w}}{\mathbf{w}^T \mathbf{w}} \right\}, \tag{2}$$

therefore the loading vectors are eigenvectors of $\mathbf{X^T X}$, where $\mathbf{X^T X}$ itself can be recognised as proportional to the covariance matrix of the dataset $\mathbf{X}$ and in this case the full principal components decomposition of $\mathbf{X}$ can be given as $\mathbf{T} = \mathbf{XW}$, where $\mathbf{W}$ is a p-by-p matrix whose columns are the eigenvectors of $\mathbf{X^T X}$ (*e.g.* Abdi and Williams, 2010).

3. Application of the PCA

In order to perform a PCA, a dense filled parametric space is required; hence from the initial sample of the 314 SEP events (Papaioannou *et al.*, 2016), a total of 126 SEP events presenting complete information with respect to all SFs and CME parameters, which in turn were treated as the variables for the PCA, were chosen. This analysis transforms the input vectors (here, each SEP event is a vector in six dimensions corresponding to the six variables extracted from the database, shown in the Appendix, Table A1) into a set of orthogonal components. The inputs of the analysis were: a) the logarithm of the peak flare flux, (log$SXRs$), b) the longitude of the associated flare, (lon), c) the flare rise time, (RT), d) the flare duration time, (DT), e) the velocity of the CME, (u) and f) the size of the CME, (s).

In our analysis we used the weighted principal component analysis (Abdi and Williams, 2010; Jolliffe, 2002). First, we centred our variables so that the mean of each column of the matrix $\mathbf{X}$ is equal to zero. Then, we used as weights the inverse variable variances while performing the PCA. Despite the fact that the PCA is a mathematically optimal method, it is sensitive to outliers in the data that produce large errors, which in turn the PCA tries to avoid. In the weighted PCA, the algorithm increases robustness by assigning different weights to data, based on their estimated relevancy, therefore the contribution of the outliers is reduced. Next, we computed the principal component transformation using the singular value decomposition (SVD) of $\mathbf{X}$.

Table 1. Results of the Principal Component Analysis (PCA)

Component	Latent	Variance (%)	Cumulative (%)
PC1	2.485	41.42	41.42
PC2	1.314	21.90	63.32
PC3	0.997	16.61	79.94
PC4	0.649	10.82	90.76
PC5	0.447	7.44	98.20
PC6	0.108	1.79	100

Table 1 presents the outputs of the method. Column 1 provides the number of the component, column 2 presents the corresponding eigenvalues of the covariance matrix of the six variables of our database (*i.e.* the latent), column 3 gives the variance expressed in percentages and column 4 shows the cumulative variance, again in percentages. The first principal component PC1 explains 41.42% of the variation, with the following three components, *i.e.*, PC2, PC3, and PC4 that correspondingly explain, the 21.90%, 16.61%, and 10.82% of the variation. The first four components (*e.g.* PC1 – PC4) account for the 90.76% of the variation, while the other two components (*e.g.* PC5 and PC6) explain the remaining ∼10% of the variation.

Based on the findings presented in Table 1, Figure 1 displays the number of the principal component *versus* its corresponding eigenvalue, ordered from the largest to the smallest. This is the so-called scree plot and it depicts the explained variance as a function of the principal components.

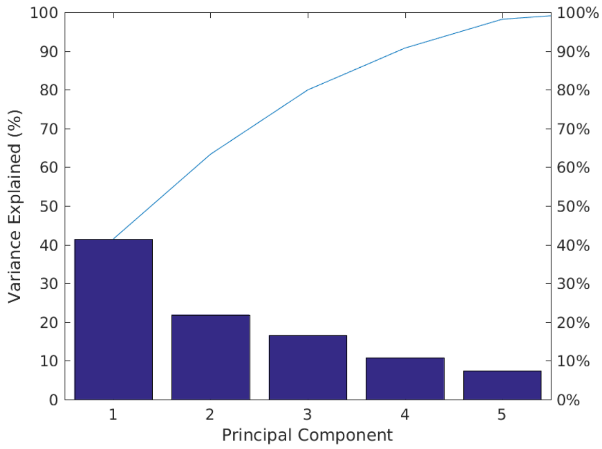

Figure 1. Scree plot of the percentual variability explained by each principal component.

Next the correlation between the first two principal components, which seem to be the dominant ones in our sample, and the original variables, called component loadings, is presented in Table 2. Column 1 provides the initial variables, columns two to seven present the calculated loadings *per* principal component. Focusing on the first two principal components, it can be seen that the highest component loading for PC1 comes from the velocity of the CME (u), of the width of the CME (w), and the logarithm of the peak flare flux (log$SXRs$), while PC2 loads on the flare duration time (DT), and the the flare rise time (RT).

Table 2. Principal component loadings

Variables	PC1	PC2	PC3	PC4	PC5	PC6
velocity of the CME (u)	0.4145	0.4478	-0.0202	0.3786	0.6955	-0.0164
width of the CME (w)	0.4474	0.3584	-0.0701	0.3938	-0.7145	-0.0293
flare duration (DT)	0.5012	-0.4940	-0.0137	-0.0024	0.0370	0.7094
flare longitude (Lon)	0.0504	0.0330	0.9972	0.0255	-0.0360	0.0086
flare rise time (RT)	0.4954	-0.5032	0.0012	-0.0678	0.0492	-0.7031
log. peak flare flux (logSXR)	0.3591	0.4156	-0.0118	-0.8345	-0.0269	0.0341

3.1. Classification of SEP Events

As stated in Section 3, it is possible to interpret the principal components in a meaningful manner and identify structures reflected in the obtained results. To this end, Figure 2 presents the score[3] plots of the SEP sample for different groups of the initial parameters and variables (from the top panel on the left, labelled with a, to the bottom panel on the left, labelled with g), as well as the loading plot (bottom panel on the right, labelled with h). The first four panels of Figure 2 focus on the variables that stem from solar flares, while the following two panels, *i.e.*, e and f display the obtained score plots on the basis of the CME characteristics. Panel a is colour coded on the basis of the position of the parent solar flare, *i.e.*, green stands for western longitudes (W20−W120), blue for central longitudes (E20-W20), and brick for eastern (E90-E20) longitudes of the SEP associated solar flares. Next, panel b is colour coded on the basis of the GOES peak photon flux with blue colour presenting C class, brick colour M class, and green colour X class solar flares. Panel c is colour coded on the basis of the solar flare rise time, with blue colour denoting gradual flares (*i.e.* rise time ≥ 13 min; Park *et al.* (2010)) and brick colour standing for impulsive solar flares (i.e. rise time < 13 min). Furthermore, panel d is colour coded with respect to the duration of the solar flare. Blue colour stands for long duration solar flares (*i.e.* those lasting ≥ 60 min), while brick colour for short duration solar flares (*i.e.* those lasting < 60 min). These four variables represent the timing, the position as well as the magnitude of the solar flares, associated with SEPs. The next two panels in Figure 2 are colour coded on the basis of the CME characteristics. Panel e presents halo (Earth directed, 360° width) CMEs in blue colour and all

[3]In the weighted PCA scores are calculated as follows: $X - mean(X)/variance(X)$

other non halo CMEs in brick colour. Panel f depicts fast CMEs ($\geq$1000 km s^{-1}) in blue and slow CMEs (<1000 km s^{-1}) in brick colour.

In addition, since the peak proton flux of each of the 126 SEP events was precalculated in our database, we distributed the events with respect to their achieved solar storm level[4]. Panel g presents the score plot of all 126 events colour coded as a function of their solar radiation scale, *e.g.* S1 in brick colour, S2 in green, S3 in purple, S4 in grey, and minor events (E> 10 MeV <10 pfu) in blue colour. This is directly comparable to panels a-f and demonstrates the effect of the different groupings (classification) on the derived peak proton flux of the SEP events in our sample.

Finally, panel h depicts all six variables of our database, represented by a vector (*e.g.* load vector), and the direction and length of the vector indicates how each variable contributes to the two principal components, *i.e.* the loading of each variable to the first two principal components as those are also presented in Table 2. In this 2D biplot (that overlays the score and the loading plot) we also include a point for each of the 3663 observations, with coordinates indicating the score of each observation for the two principal components in the plot. These points are scaled with respect to the maximum score value and the maximum coefficient length, thus only their relative locations can be determined from the biplot. Red colour stands for non-SEP entries in the database, while blue colour represents SEP related entries. The ends of the vectors represent the correlations of each variable with each component and the direction of the vectors shows that the values of the variable increase in that direction. The first principal component, on the horizontal axis, (PC1) has positive coefficients for all six variables, while the CME variables w and u, as well as the logSXR seem to load high in PC1. At the same time, the second principal component (PC2), on the vertical axis, has negative coefficients for the variables DT and RT, and positive coefficients for the remaining four variables. Inspection of Table 2 shows that PC2 significantly loads on DT and RT. The variable (lon) has the lowest contribution to the first two principal components. From panel h of Figure 2, one can see that the velocity of the CME (u), the size of the CME (s), and the logarithm of the peak flare flux (log$SXRs$) load high in PC2, while the duration of the solar flare (DT), and the the flare rise time (RT) load high in PC1. As a result, two groups can be distinguished.

A comparison of the score plots a-f to the score plot [g] identifies which SEP events will result in enhanced peak proton fluxes (at E> 10 MeV). In particular, SEP events related to fast and halo CMEs (panels e, f), as well as solar flares of significant importance (> M class - panel b) lead to significant peak proton fluxes, categorized as S4, S3, and S2 solar radiation storms (panel g). On the other hand, slow and non–halo CMEs associated with small, in magnitude, solar flares (C class), result to minor or S1 solar radiation storms. Furthermore, impulsive and short duration solar flares (panels c, d) are mostly situated on the western part of the visible solar disk (panel a), are associated to significant solar flares (M and X class) and result to enhanced radiation storms (panel g).

[4] http://www.swpc.noaa.gov/noaa-scales-explanation

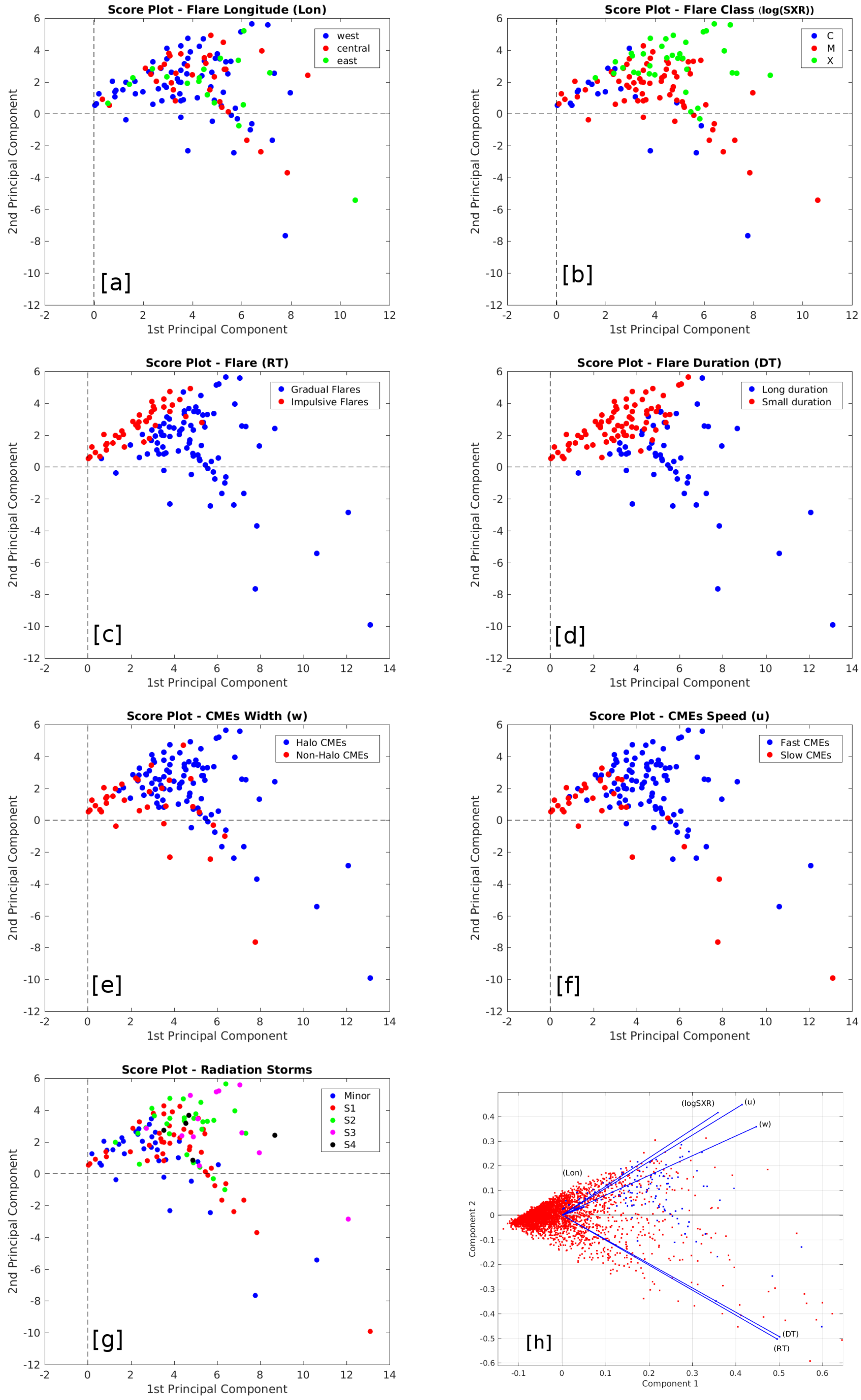

Figure 2. Results of the PCA. From top to bottom, seven score plots, colour coded on the basis of different groupings of the variables (see text for details), while the bottom panel on the right depicts a 2D biplot.

Finally, gradual and long duration solar flares are attributed mostly to M class flares, with minor or S1 solar radiation storms, being prevalent.

4. SEPs Short-Term Forecasting (Nowcasting) Based on PCA

As a next step, an attempt was made to identify whether the results from the multi-variable PCA can be used to quantify the occurrence (or not) of an SEP event. This is because, as denoted above (see Section 3) the parametric space of the two principal components may lead to a dichotomous separation between SEP events and non-SEP events. To this end, we further investigated the nowcasting capabilities of the PCA parametric space with the application of the logistic regression method (Garcia, 2004; Laurenza *et al.*, 2009; Winter and Ledbetter, 2015). Our results are summarized in the following:

4.1. Application of the Logistic Regression

At this point we applied the logistic regression analysis, a statistical method in which there are one or more independent variables (the PCA components in our case) that determine an outcome which is the dependent variable (Hosmer Jr, Lemeshow, and Sturdivant, 2013). The outcome is a binary or dichotomous variable, *i.e.* there are only two possible outcomes, 1 (SEP events in our case, TRUE or success) or 0 (non-SEP events in our case, FALSE or failure).

The main purpose of the logistic regression analysis is to find the best fitting model in order to describe the relationship between the dichotomous characteristic of interest (dependent variable, response or outcome) and a set of independent (predictor or explanatory variable) variables, which can be discrete and/or continuous (Hosmer Jr, Lemeshow, and Sturdivant, 2013). Rather than choosing parameters that minimize the sum of squared errors (like in the ordinary regression), the logistic regression analysis estimates the parameters that maximize the likelihood of observing the sample values. The application of this method generates the coefficients of a sigmoidal function to predict a logit transformation (*i.e.* the inverse of the sigmoidal "logistic" function) of the probability (Harrell, 2001).

In detail, we considered a generalized "logistic" function to model the SEP occurrence probability as a function of the explanatory variables, which in our analysis will be the PCA components. The "logistic" function is defined as:

$$h_\theta(g(x)) = \frac{1}{1 + e^{-\theta^T g(x)}} \tag{3}$$

and it is parametrized by θ, which are the coefficients of the function, and $g(x)$ which is a function of the explanatory variables x_i. In connection to the principal components, the explanatory variables x_i can be defined either as vector-matrix or as an n-dimensional matrix of any linear or non-linear relation between the principal components (PC1, PC2, ...). In particular, in the one-parametric linear logistic regression case x_i is a vector matrix, 1_i is the unity matrix and $g(x)$= $(1_i,\ x_i)$ = (1,PC1_i) or (1,PC2_i) or (1,PC1_i+PC2_i) and the product $\theta^T g(x) =$

$\theta_0 + \theta_1 x_i$, where $x_1, x_2, ..., x_i$ are the explanatory variables defined above. In the multivariate case (*e.g.* multiple logistic regression), x_i is a matrix and $g(x) = (1_i, x_i^j) = (1_i, x_i^1, x_i^2, ...)$, where each column-vector x^j can be defined in any linear or non-linear relation between the PCs. Therefore, in the multiple logistic regression the product $\theta^T x_i^j = \theta_0 + \theta_1 x_i^1 + \theta_2 x_i^2 + ...$ Moreover, the independent (explanatory) variables can be even the power terms or some other non-linear transformations of the original independent variables (interaction terms), for example the simplest case of multiple non-linear logistic regression with two explanatory variables will have $\theta^T g(x) = \theta_0 + \theta_1 x_i^1 + \theta_2 x_i^2 + \theta_3 x_i^1 x_i^2$. In our analysis we will apply different logistic regression probabilistic models based on the selection of the function $g(x)$, to estimate their accuracy and their categorical scoring in every case.

To estimate the coefficients, θ, of the "logistic" function we used the principle of maximum likelihood, therefore we need to minimize the negative log likelihood function (*i.e.* the cost function), given the current training set (Shevade and Keerthi, 2003). For the logistic regression, the cost function is defined as:

$$J(\theta) = -\frac{1}{m} \sum_{n=1}^{m} y^{(i)} log(h(x^{(i)})) + (1 - y^{(i)}) log(1 - h(x^{(i)})), \tag{4}$$

which is a convex cost function that can be derived from statistics using the principle of maximum likelihood estimation (Govan, 2006). To minimize the logistic regression cost function we use an advanced cost minimization algorithm which is based on the BFGS (Broyden-Fletcher-Goldfarb-Shanno) Quasi-Newton method (Head and Zerner, 1985; Schraudolph, Yu, and Günter, 2007).

5. A Possible Index for the Prognosis of SEP Events

5.1. Logistic Regression With One Predictor or Explanatory Variable

In this scheme we produced an index (I) from the estimated principal components of the flare and CME parameters of Section 3. Our purpose was to determine if such an index could be used for the forecasting of the occurrence of SEPs. In order to effectively use the new index for SEP forecasting, there should be an apparent separation between the two categories, *i.e.* the non-SEP events and the SEP ones, based on the distribution characteristics (mean value, variance) of each case. With the use of box-plots we show in Figure 3 the distributions of the first three principal components (*e.g.* PC1, PC2, and PC3) for the two separate categories (responses). For the first and the second principal component (PC1 and PC2) the SEP events are clearly separated from the non-SEP ones, while for the third component there is no apparent separation. From the results of the Figure 3 (panel on the left) it is clear that one may attempt to make use of the first two principal components as a new forecasting index. We started our analysis with the first principal component (PC1) of the PCA and we defined the index (I) as follows:

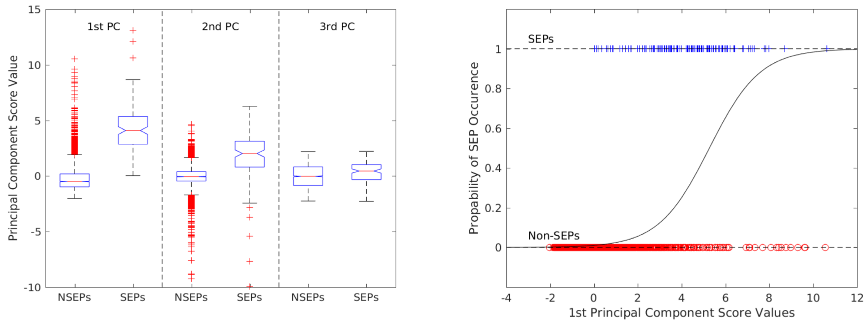

Figure 3. Box plots of the principal component score values for the non-SEP and the SEP events separately. The red line inside the box indicates the median of the distributions, the bottom and top edges of the box indicate the first and third quartiles respectively (*i.e.* 25th and 75th percentile), and the outermost lines indicate the maximum and minimum values of the distribution without the outliers which are depicted with the red crosses (panel on the left hand side). The resulting fitting from the logistic regression (panel on the right hand side), see text for details.

$$I = \mathrm{PC1} = \mathrm{A}_1 \cdot \log SXR + \mathrm{A}_2 \cdot long + \mathrm{A}_3 \cdot RT + \mathrm{A}_4 \cdot DT + \mathrm{A}_5 \cdot u + \mathrm{A}_6 \cdot w \quad (5)$$

The coefficients $A_1, ..., A_6$ are the loadings of the first principal component (PC1) that have been estimated from the PCA (see column 2 of Table 2), so that most of the variance (*i.e.* 41.42 %) of the initial observations, *i.e.* the SF and CME parameters (see Table 1 and Figure 1), is taken into account. Next, we applied the logistic regression method with one predictor (explanatory) variable in order to identify the probability of SEP occurrence as a function of the new index I. From the logistic regression we estimated the parameter θ that best fits to the response variable, *i.e.* the two categories: SEP events or non-SEP events (see Equations 3 and 4). The resulting fitting from the logistic regression is depicted in Figure 3 (panel on the right). In particular, this panel presents the logistic regression curve that depicts the probability of having an SEP (or non-SEP) event as a function of the index I which for this example was selected to be the PC1. The θ parameter controls the characteristics of the logistic regression curve and the blue and the red points are the actual observations (SEPs or non-SEPs) that has to be fitted in the probabilistic sense with the logistic function.

From this analysis we found that the cost function reaches a minimum for $\theta = [-4.553, 0.865]$ and for a probability threshold of 50%, which is expressed as an index value of $I = 5.264$, 27.0% of the SEP events lie above and 99.2% of the non-SEP ones lie below this index value. These measures can be better realized by constructing a confusion matrix (a special kind of contingency table; Anastasiadis *et al.* (2017); Davis and Goadrich (2006) for a probability threshold of 50%, therefore, we have 34 *true positive* (TP, a) predictions, 27 *false positive* (FP, b) predictions, 3510 *true negative* (TN, d) predictions, and 92 *false negative* (FN, c) predictions. From the above values we calculated the probability of detection (POD, $a/a+c$) and the probability of a false alarm (PFA) or false-alarm rate (FAR, $b/a+b$) (Balch, 2008; Anastasiadis *et al.*, 2017). We found that with the use of the first principal component as a predictor variable we have a relatively high false-alarm rate, $FAR = 44.3\%$ (27/61) and the probability of detection was low, $POD = 27.0\%$ (34/126).

We additionally used as an index the second principal component (PC2) and a linear combination of PC1 and PC2 (*i.e.* PC1+PC2), and we applied again the logistic regression method with one predictor (explanatory) variable, in order to investigate if the predictions qualitatively change. From the logistic regression we found that using I = PC2 the overall accuracy of the scheme significantly dropped, resulting in a POD = 15.1%. It seems that the use of the second component as an index cannot effectively separate our sample into the two categories. As a next step, using I = PC1+PC2 we found results improved to those obtained for I =PC1. The POD was 43.7% (55/126) and the FAR was 31.25% (25/80). The probability of detection improved significantly and, at the same time, we gained a relatively lower FAR.

5.2. Multivariate Logistic Regression

As a next step we applied a multivariate logistic regression (Tabachnick and Fidell, 2007) to examine the SEP occurrence probability as a function of an index with multiple explanatory variables. In this case the index was treated as a multidimensional array comprising of the principal components of the PCA. We started with the simplest case which is the 2D logistic regression of the first two principal components. In this case the index is defined as I = [PC1, PC2], where PC1 and PC2 are arrays of the first and the second principal component score values, respectively, and their dimension is 1 × N where N is the length of our dataset (*i.e.* 126 SEPs + 3537 non-SEPs = 3663 records), therefore, I is a 2 × N dimensional array. The application of the multivariate logistic regression is based on the method presented in Section 4.1.

In the left hand side of Figure 4 we show a scatter plot of the two principal components. Red circles depict the non-SEP events and blue crosses the SEP events. From the characteristics of this figure it is clear that the use of the two principal components in a multivariate regression can effectively separate the events into the two categories of non-SEPs and SEPs. Although there is significant scatter from the perfect dichotomous prediction case, SEP events tend to be grouped in a region that can be visually separated from the region where non-SEPs appear. We performed the multivariate logistic regression in the first two principal components and we found that the cost function reaches a minimum for $\theta = [-5.555, 1.042, 0.719]$. In Figure 4 we show with a straight line the resulting decision boundary for a probability threshold of 50%.

From the results of the multivariate logistic regression we constructed the confusion matrix and we found 69 TP, 57 FN, 3507 TN, and 30 FP predictions. From the confusion matrix we also calculated the POD and the FAR of this scheme for a probability threshold of 50%. We found that the POD of this scheme was 54.8%(69/126) and the FAR was 30.3%(30/99) which are both significantly better than the POD and FAR that we estimated with the logistic regression of the 1D index in the previous section.

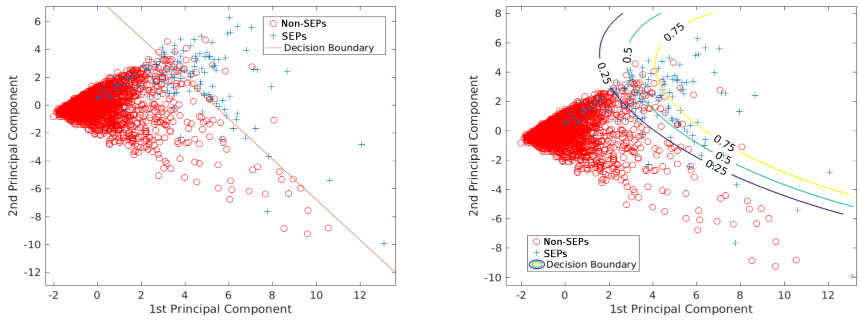

Figure 4. Scatter plot of the SEP (blue crosses) and non-SEP (red circles) events, as they map on the projected space of PC1 and PC2. The decision boundary for a $p_{th} = 50\%$ of the $I^{(2)}$ scheme is depicted in the left hand side panel, while three decision boundaries for $p_{th} =$ 25%, 50% and 75% of the $I^{(2+O^2)}$ scheme are depicted in the right hand side panel. See text for details.

Table 3. Summary of Categorical Scores *per* scheme

index	Form (Scheme)	POD (%)	FAR (%)	HSS
$I^{(1)}$	[PC1]	26.98	44.26	0.3490
$I^{(2)}$	[PC1, PC2]	54.76	30.30	0.6013
$I^{(3)}$	[PC1, PC2, PC3]	55.56	28.57	0.6134
$I^{(4)}$	[PC1, PC2, PC3, PC4]	53.97	29.17	0.6007
$I^{(5)}$	[PC1, ..., PC5]	53.17	29.47	0.5943
$I^{(6)}$	[PC1, ..., PC6]	53.17	28.72	0.5973
$I^{(2+O^2)}$	$[\mathrm{PC1}, \mathrm{PC2}, \mathrm{PC1}^2, \mathrm{PC2}^2, \mathrm{PC1} \cdot \mathrm{PC2}]$	56.35	31.07	0.6080
$I^{(3+O^2)}$	$[\mathrm{PC1}, \mathrm{PC2}, \mathrm{PC3}, \mathrm{PC1}^2, \mathrm{PC2}^2, \mathrm{PC3}^2, \mathrm{PC1} \cdot \mathrm{PC2}]$	58.73	24.49	0.6502

Further, we extended our analysis using different combinations of the principal components for the construction of the index as a matrix. We started by adding to the matrix $I^{(2)}$ = [PC1,PC2] one component at the time until we included all six components (*e.g.* $I^{(6)}$ = [PC1,PC2,PC3,PC4,PC5,PC6]). In every case we performed multivariate logistic regression and we calculated the POD and FAR for every new index to examine its performance. The results for the derived POD and FAR are presented in Table 3. From this analysis it seems that the resulting POD and FAR do not change significantly with the addition of more components to the index matrix. The best POD is obtained for $I^{(2)}$, while the best FAR is obtained for $I^{(3)}$. The optimal score for each index can be traced using the Heidke skill score (HSS), which is a measure of skill in forecasts and quantifies the ability of achieving correct predictions with respect to chance. For a probability threshold 50% we found that the best optimal HSS is obtained for $I^{(3)}$, while for $I^{(2)}$ we have the next best score (see Table 3).

5.3. Multivariate Logistic Regression with Interaction Terms

In this part of our analysis we performed a logistic regression with the inclusion of interaction terms in the index matrix. The interaction terms are usually either square (or higher order) values of the initial explanatory variables (*i.e.* [PC1^2,PC2^2, ...]) or products of the explanatory variables (*i.e.* [PC1·PC2,PC2·PC3, ...]). With this method the decision boundary is a non-linear function and its parametric form will depend on the selection of the interaction terms. For example in Figure 4 (panel on the left), where no interaction terms are included in the model, the decision boundary is a straight line (PC2 $= a + b$PC1) which separates SEPs from non-SEPs. Higher order terms would lead to complex boundaries with higher order parametric forms.

In addition, we examined if the inclusion of the interaction terms into the logistic regression analysis scheme leads to an improvement of the prediction accuracy of our model. We started from the simplest case of $I^{(2)} =$ [PC1,PC2] and we added interaction terms in the form of $I^{(O^2)} =$ [PC1^2,PC2^2,PC1·PC2]. The new index matrix becomes $I^{(2+O^2)} =$ [PC1,PC2,PC1^2,PC2^2,PC1·PC2]. We found that the cost function becomes minimum for $\theta = [-6.043, 1.379, 1.188, -0.041, -0.058, -0.105]$. Additionally, we found 71 TP, 55 FN, 3505 TN, and 32 FP predictions which yield a POD of 56.35% (71/126) and a FAR of 31.07% (32/103) (for a threshold set at 50%). The HSS was found to be 0.608 therefore the performance of this scheme seems to be better (see Table 3). Figure 4 (panel on the right) illustrates the resulting decision boundaries for three different probability thresholds (*i.e.* p_{th} = 25%, 50%, and 75%) for this scheme. It seems that the inclusion of the non-linear terms in the index matrix, which also results in a non-linear decision boundary, improves the overall performance of our method.

We further extended the above method by considering more principal components in the index matrix and by adding the corresponding interaction terms. Since the complexity of the method increases significantly with the addition of new components, we limited our analysis up to the fourth principal component. From this analysis we found that after the inclusion of the fourth component and its interaction terms, in the model, the performance remained almost constant.

6. Categorical Scores

The schemes with the best skill score were $I^{(3)}$ and $I^{(3+O^2)}$ (see Table 3). As a result, we calculated their categorical measures as a function of the probability threshold. That is we treated p_{th} as an independent parameter (not set to 50%, as this was the case in Section 5) ranging between 0.0 to 1.0 with a step of 0.1. For both schemes we then constructed the performance categorical quality measures POD, FAR, and HSS, which are considered as functions of p_{th} (Laurenza *et al.*, 2009; Anastasiadis *et al.*, 2017).

Figure 5 depicts the categorical quality measures for $I^{(3)}$ (panel on the left) and $I^{(3+O^2)}$ (panel on the right) *versus* the p_{th} level. POD (blue colour line), FAR (brick colour line), and HSS (orange colour line) are presented in each of the two panels. Both POD and FAR are significantly high and tend to decrease when

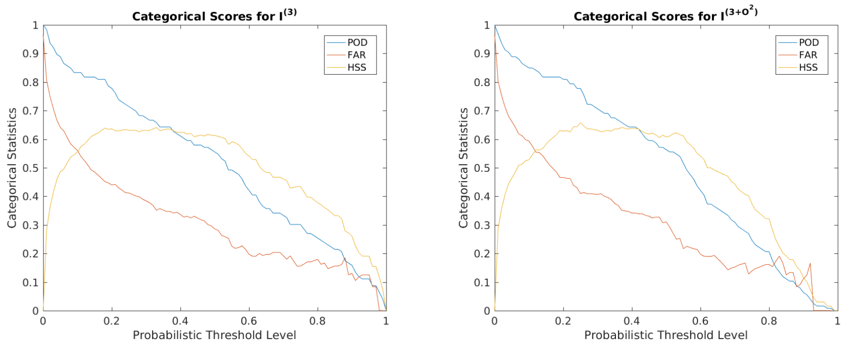

Figure 5. Categorical Scores (POD, FAR, HSS; see text for details) for $I^{(3)}$ and $I^{(3+O^2)}$.

p_{th} increases. The optimal skill score for p_{th} is an actual settlement in order to achieve maximum POD, minimum FAR, and optimized HSS. For both schemes, the optimal skill score is achieved at a range of p_{th} from 25% to 40%. The optimal HSS is observed at $p_{th} = 0.33$ (HSS=0.6411) for $I^{(3)}$ and at $p_{th} = 0.25$ (HSS=0.6579) for $I^{(3+O^2)}$. In turn, this results in a POD = 65.87% and a FAR = 35.16% as well as a POD = 77.78% and a FAR = 40.96%, respectively.

7. Discussion and Conclusions

We analysed 126 SEP events and 3537 non-SEP events with complete solar associations expressed in six variables (*i.e.* a) the logarithm of the peak flare flux (log$SXRs$), b) the longitude of the associated flare (lon), c) the flare rise time (RT), d) the flare duration (DT), e) the velocity of the CME (u), and f) the size of the CME (s)), occurring in 1997 − 2013.

Next, we applied a Principal Component Analysis (PCA) to the SEP events of our sample and showed that significant radiation storms, categorized as S4, S3 and S2, are related to fast and halo CMEs, as well as SFs of class larger than M. PCA, also, showed that impulsive and short duration, strong (M and X class) SFs mostly situated on the west part of the visible solar disk also result to enhanced radiation storms, as illustrated in the different panels of Figure 2. These results are in agreement with and, actually, summarize earlier independent studies (*e.g.* Belov *et al.*, 2005; Cane, Richardson, and Von Rosenvinge, 2010; Huang, Wang, and Li, 2012; Park and Moon, 2014; Papaioannou *et al.*, 2016; Belov, 2017; Paassilta *et al.*, 2017) but contradict the results presented by Park, Moon, and Lee (2017) who concluded that the longitudinal separation angle is the most important parameter with respect to the SEP peak flux.

Furthermore, using the outputs of the PCA, a new index (I) was introduced and tested with respect to its predictive capabilities. It was demonstrated that it actually holds prognosis potential for SEP events. Employing the logistic regression analysis, we introduced several different schemes for the I index, starting from one predictor or explanatory variable, going to multiple explanatory variables, treating I as a multidimensional array. We found that the statistical classification of SEP events *versus* non-SEP ones, based on the PCA and the related solar variables, for a threshold p_{th}=50% leads to a FAR of 24.49 % while correctly predicting 58.73 % of solar events as SEP *versus* non-SEP events (see Section 5).

As a final step, when we treated the probabilistic threshold as an independent variable ranging from 0.0 to 1.0 and calculated the categorical measures (POD, FAR, HSS) we showed that the optimal skill score was achieved at a range of p_{th} from 25% to 40% for two configuration of the index I, *i.e.* $I^{(3)}$ = [PC1,PC2,PC3] and $I^{(3+O^2)}$ = [PC1,PC2,PC3,PC1^2,PC2^2,PC3^2,PC1·PC2]. In particular, for $I^{(3)}$ this was achieved at p_{th} = 0.33 (HSS=0.6411) with POD = 65.87% and a FAR = 35.16%. At the same time for $I^{(3+O^2)}$ the relevant outputs were p_{th} = 0.25 (HSS=0.6579), with POD = 77.78% and a FAR = 40.96%. These results show that when the PCA is applied to SEP events and their parent solar sources, as defined by a multi-variable data grid parametrised from SFs (longitude, maximum soft X-ray flux, rise time, duration) and CMEs (velocity, width) characteristics, together with the logistic regression analysis, it is possible to predict the occurrence (or not) of SEP events. Our results are comparable to the derived POD and FAR of the Empirical Model for Solar Proton Events Real Time Alert (ESPERTA) concept that utilized a logistic regression scheme on, basically, two parameters: (i) the SXR fluence and (ii) the radio fluence at ≈ 1 MHz for three different longitudinal bands (Alberti *et al.*, 2017). This highlights the fact that the outcome of any treatment (*e.g.* PCA with logistic regression or logistic regression alone) depends on which solar observables (variables) are used.

Furthermore, it is noteworthy, that most of the SEP prediction concepts that rely on empirical or semi-empirical relations are in need of solar observables *i.e.* precursor data, which in turn are used as variables (inputs). Therefore, if no identification of an SF or a CME is available (for example, if a behind the limb SF is taking place) and an SEP does occurs, such an event will be missed (not forecasted). At the same time, the work from Posner (2007) has proven the concept of short-term forecasting of the appearance and intensity of solar ion events utilizing *in situ* relativistic electron recordings, making use of the higher speed of these electrons propagating from the Sun to 1 AU.

Our results should be considered as a first step towards an integrated SEP event prognosis. Given the current wealth of observations at hand and the association of SEP events to both SFs and CMEs, multi–variate methods may hold a key for future advances in the field. It has already been noted by Winter and Ledbetter (2015) that when applying PCA to type II bursts it was possible to achieve a POD=62% and an FAR=21% (their Table 8). Further work is necessary in order to refine the proposed index I, in terms of the variables used in the PCA. For example, the duration of the SF (DT), as well as the width of the CME (s) are particularly uncertain parameters. Furthermore, it is desirable to go beyond the nowcasting of the occurrence (or not) of SEP events and try to quantify the expected impact in terms of the expected radiation storm level.

Acknowledgments AP would like to acknowledge support from a post-doctoral IKY scholarship funded by the action "Supporting post-doctoral researchers" from the resources of the b.p. “Human Resources Development Education and Lifelong Learning” with Priority Axes 6, 8, 9 and co-funded by the European Social Fund and the Greek government. AA would further like to acknowledge the “SPECS: Solar Particle Events and foreCasting Studies research grant of the National Observatory of Athens. MP and RV acknowledge the funding from the Academy of Finland (decision 267186). Research conducted by MP and RV was further supported by ESA contract 4000120480/17/NL/LF/hh. The authors would further like to thank the anonymous referee for constructive comments that helped to improve the initial manuscript.

Disclosure of Potential Conflict of Interest The authors declare that they have no conflict of interest

References

Abdi, H., Williams, L.J.: 2010, *Wiley interdisciplinary reviews: computational statistics* **2**(4), 433. DOI. 10.1002/wics.101.

Alberti, T., Laurenza, M., Cliver, E., Storini, M., Consolini, G., Lepreti, F.: 2017, *Astrophys J* **838**(1), 59. DOI. 10.3847/1538-4357/aa5cb8.

Anastasiadis, A., Papaioannou, A., Sandberg, I., Georgoulis, M., Tziotziou, K., Kouloumvakos, A., Jiggens, P.: 2017, *Solar Phys* **292**(9), 134. DOI. 10.1007/s11207-017-1163-7.

Anastasiadis, A.: 2002, *J Atmos. Sol-Terr Phy* **64**(5), 481. DOI. 10.1016/S1364-6826(02)00003-2.

Balch, C.C.: 1999, *Radiat Meas* **30**(3), 231. DOI. 10.1016/S1350-4487(99)00052-9.

Balch, C.C.: 2008, *Space Weather* **6**(1). DOI. 10.1029/2007SW000337.

Belov, A.: 2009, *Adv. Space Res.* **43**(4), 467. DOI. 10.1016/j.asr.2008.08.011.

Belov, A., Garcia, H., Kurt, V., Mavromichalaki, H., Gerontidou, M.: 2005, *Solar Phys* **229**(1), 135. DOI. 10.1007/s11207-005-4721-3.

Belov, A.: 2017, *Geomagn Aeronomy* **57**(6), 727. DOI. 10.1134/S0016793217060020.

Cane, H., Lario, D.: 2006, *Space Sci Rev* **123**(1-3), 45. DOI.

Cane, H., Richardson, I., Von Rosenvinge, T.: 2010, *J. Geophys Res* **115**(A8). DOI. 10.1029/2009JA014848.

Chancellor, J.C., Scott, G.B., Sutton, J.P.: 2014, *Life* **4**(3), 491.

Davis, J., Goadrich, M.: 2006, The relationship between precision-recall and roc curves, 233. 1-59593-383-2. DOI. 10.1145/1143844.1143874.

Dierckxsens, M., Tziotziou, K., Dalla, S., Patsou, I., Marsh, M., Crosby, N., Malandraki, O., Tsiropoula, G.: 2015, *Solar Phys* **290**(3), 841. DOI. 10.1007/s11207-014-0641-4.

Dresing, N., Gómez-Herrero, R., Klassen, A., Heber, B., Kartavykh, Y., Dröge, W.: 2012, *Solar Phys* **281**(1), 281. DOI. 10.1007/s11207-012-0049-y.

Dröge, W., Kartavykh, Y., Klecker, B., Kovaltsov, G.: 2010, *Astrophys J* **709**(2), 912. DOI. 10.1088/0004-637X/709/2/912.

Engell, A., Falconer, D., Schuh, M., Loomis, J., Bissett, D.: 2017, *Space Weather* **15**(10), 1321. DOI. 10.1002/2017SW001660.

Garcia, H.: 2004, *Space Weather* **2**(6). DOI. 10.1029/2003SW000035.

Gómez-Herrero, R., Dresing, N., Klassen, A., Heber, B., Lario, D., Agueda, N., Malandraki, O., Blanco, J., Rodríguez-Pacheco, J., Banjac, S.: 2015, *Astrophys J* **799**(1), 55. DOI. 10.1088/0004-637X/799/1/55.

Govan, A.: 2006, In: *North Carolina State University, SAMSI NDHS, Undergraduate workshop.* https://projects.ncsu.edu/crsc/events/ugw06/presentations/aygovan/OptimizationUW06.pdf.

Harrell, F.E.: 2001, Ordinal logistic regression, Springer, New York, NY, 331. 978-1-4757-3462-1. DOI. 10.1007/978-1-4757-3462-1_13.

Head, J.D., Zerner, M.C.: 1985, *Chem Phys Lett* **122**(3), 264. DOI. 10.1016/0009-2614(85)80574-1.

Hosmer Jr, D.W., Lemeshow, S., Sturdivant, R.X.: 2013, *Applied logistic regression* **398**, John Wiley & Sons, Hoboken, New Jersey. ISBN 978-0-470-58247-3.

Huang, X., Wang, H.-N., Li, L.-P.: 2012, *Res Astron Astrophys* **12**(3), 313. DOI. 10.1088/1674-4527/12/3/007.

Iucci, N., Levitin, A., Belov, A., Eroshenko, E., Ptitsyna, N., Villoresi, G., Chizhenkov, G., Dorman, L., Gromova, L., Parisi, M., *et al.*: 2005, *Space Weather* **3**(1). DOI. 10.1029/2003SW000056.

Jolliffe, I.: 2002, *Principal component analysis*, Springer, New York. ISBN 978-0-387-22440-4. DOI. 10.1007/b98835.

Kahler, S.: 2001, *J. Geophys Res* **106**(A10), 20947. DOI. 10.1029/2000JA002231.

Kocharov, L., Torsti, J.: 2002, *Solar Phys* **207**(1), 149. DOI. 10.1023/A:1015540311183.

Kouloumvakos, A., Patsourakos, S., Nindos, A., Vourlidas, A., Anastasiadis, A., Hillaris, A., Sandberg, I.: 2016, *Astrophys J* **821**(1), 31. DOI. 10.3847/0004-637X/821/1/31.

Kurt, V., Belov, A., Mavromichalaki, H., Gerontidou, M.: 2004, Statistical analysis of solar proton events. **22**(6), 2255. DOI. 10.5194/angeo-22-2255-2004.

Lario, D., Kwon, R.-Y., Vourlidas, A., Raouafi, N., Haggerty, D., Ho, G., Anderson, B., Papaioannou, A., Gómez-Herrero, R., Dresing, N., *et al.*: 2016, *Astrophys J* **819**(1), 72. DOI. 10.3847/0004-637X/819/1/72.

Lario, D., Kwon, R.-Y., Richardson, I.G., Raouafi, N.E., Thompson, B., Von Rosenvinge, T.T., Mays, M.L., Mäkelä, P.A., Xie, H., Bain, H., *et al.*: 2017, *Astrophys J* **838**(1), 51. DOI. 10.3847/1538-4357/aa63e4.

Laurenza, M., Cliver, E., Hewitt, J., Storini, M., Ling, A., Balch, C., Kaiser, M.: 2009, *Space Weather* **7**(4). DOI. 10.1029/2007SW000379.

Lim, M.: 2002, *Occup Environ Med* **59**(7), 428. DOI. 10.1136/oem.59.7.428.

Mikaelian, T.: 2009, *arXiv preprint arXiv:0906.3884*. https://arxiv.org/abs/0906.3884.

Mishev, A.: 2014, *Adv Space Res* **54**(3), 528. DOI. 10.1016/j.asr.2013.10.010.

Miteva, R., Samwel, S.W., Krupar, V.: 2017, In: Georgieva, K., Kirov, B., Danov, D. (eds.) *Proceedings of the Ninth Workshop on Solar Influences on the Magnetosphere, Ionosphere and Atmosphere* **30**, 19.

Núñez, M.: 2011, *Space Weather* **9**(7). 10.1029/2010SW000640.

Paassilta, M., Raukunen, O., Vainio, R., Valtonen, E., Papaioannou, A., Siipola, R., Riihonen, E., Dierckxsens, M., Crosby, N., Malandraki, O., *et al.*: 2017, *J. Space Weather Space Clim.* **7**, A14. DOI. 10.1051/swsc/2017013.

Papaioannou, A., Anastasiadis, A., Sandberg, I., Georgoulis, M., Tsiropoula, G., Tziotziou, K., Jiggens, P., Hilgers, A.: 2015, In: *Journal of Physics: Conference Series* **632**, 012075. IOP Publishing. DOI. 10.1088/1742-6596/632/1/012075.

Papaioannou, A., Sandberg, I., Anastasiadis, A., Kouloumvakos, A., Georgoulis, M.K., Tziotziou, K., Tsiropoula, G., Jiggens, P., Hilgers, A.: 2016, *J. Space Weather Space Clim.* **6**, A42. DOI. 10.1051/swsc/2016035.

Park, J., Moon, Y.-J.: 2014, *J. Geophys. Res* **119**(12), 9456. DOI. 10.1002/2014JA020272.

Park, J., Moon, Y.-J., Lee, H.: 2017, *Astrophys J* **844**(1), 17. DOI. 10.3847/1538-4357/aa794a.

Park, J., Moon, Y.-J., Lee, D., Youn, S.: 2010, *J. Geophys. Res* **115**(A10). DOI. 10.1029/2010JA015330.

Posner, A.: 2007, *Space Weather* **5**(5). DOI. 10.1029/2006SW000268.

Reames, D.V.: 1999, *Space Sci Rev* **90**(3-4), 413. DOI. 10.1023/A:1005105831781.

Reames, D.V.: 2013, *Space Sci Rev* **175**(1-4), 53. DOI. 10.1007/s11214-013-9958-9.

Reames, D.V.: 2017, *Solar energetic particles*, *Lecture Notes in Physics*, Springer, Berlin. ISBN 978-3-319-50870-2. DOI. 10.1007/978-3-319-50871-9.

Rouillard, A., Sheeley, N., Tylka, A., Vourlidas, A., Ng, C., Rakowski, C., Cohen, C., Mewaldt, R., Mason, G., Reames, D., *et al.*: 2012, *Astrophys J* **752**(1), 44. DOI. 10.1088/0004-637X/752/1/44.

Schraudolph, N.N., Yu, J., Günter, S.: 2007, In: *Artificial Intelligence and Statistics*, 436.

Shevade, S.K., Keerthi, S.S.: 2003, *Bioinformatics* **19**(17), 2246. DOI. 10.1093/bioinformatics/btg308.

Shlens, J.: 2014, *arXiv preprint arXiv:1404.1100*.

Smart, D., Shea, M.: 1989, *Adv Space Res* **9**(10), 281. DOI. 10.1016/0273-1177(89)90450-X.

Souvatzoglou, G., Papaioannou, A., Mavromichalaki, H., Dimitroulakos, J., Sarlanis, C.: 2014, *Space Weather* **12**(11), 633. DOI. 10.1002/2014SW001102.

Tabachnick, B.G., Fidell, L.S.: 2007, *Using multivariate statistics*, Allyn & Bacon/Pearson Education, Needham Heights, MA. ISBN 978-0-205-45938-2.

Tobiska, W.K., Atwell, W., Beck, P., Benton, E., Copeland, K., Dyer, C., Gersey, B., Getley, I., Hands, A., Holland, M., *et al.*: 2015, *Space Weather* **13**(4), 202. DOI. 10.1002/2015SW001169.

Trottet, G., Samwel, S., Klein, K.-L., de Wit, T.D., Miteva, R.: 2014, *Solar Phys*, 1. DOI. 10.1007/s11207-014-0628-1.

Turner, R.E.: 2006, In: Gopalswamy, N., Mewaldt, R.A., Torsti, J. (eds.) *Solar Eruptions and Energetic Particles Geophysical Monograph Series, American Geophysical Union* **165**, Wiley Online Library, Hoboken, New Jersey, 367. Chap. 33. ISBN 978-0-875-90430-6. DOI. 10.1029/GM165.

Wiedenbeck, M., Mason, G., Cohen, C., Nitta, N., Gómez-Herrero, R., Haggerty, D.: 2012, *Astrophys J* **762**(1), 54. DOI. 10.1088/0004-637X/762/1/54.

Winter, L., Ledbetter, K.: 2015, *Astrophys J* **809**(1), 105. DOI. 10.1088/0004-637X/809/1/105.

Appendix

Table A1. The 126 SEP events employed in the PCA. Column 1 provides the date of the related solar flare, column 2 shows the peak time of the SF, columns 3 to 8 provide the six variables used in our analysis, namely: CME width (s), and velocity (u), logarithm of the solar flare (SF) magnitude ($\log SXRs$), SF longitude (lon), duration (DT), and rise time (RT). Columns 9 and 10 give the SEP nowcasting results (where Hit and Miss refer to SEPs correctly predicted and SEPs that were not predicted) for $I^{(3)}$ and $I^{(3+O^2)}$, respectively.

SXR Date	SXR Start Time (hh:mm)	CME width, s (^{o})	CME Velocity, u (km s^{-1})	Logarithm of SXRs $\log SXRs$	Solar Flare position, Lon (^{o})	Solar Flare Duration, DT (min)	Solar Flare Rise time, RT (min)	SEP Forecast Result	SEP Forecast Result
1997.11.04	05:52	360	785	-3.677780725	33	10	6	Miss	Hit
1997.11.06	11:49	360	1556	-3.045757491	63	12	6	Hit	Hit
1998.04.20	09:38	243	1863	-4.853871972	90	100	43	Hit	Hit
1998.04.29	16:06	360	1374	-4.167491075	-20	53	31	Hit	Hit
1998.05.02	13:31	360	938	-3.958607305	15	20	11	Miss	Miss
1998.05.06	07:58	248	792	-3.568636228	65	22	11	Miss	Miss
1998.05.09	03:04	178	2331	-4.113509286	102	51	36	Hit	Hit
1998.06.16	18:03	100	1484	-5	115	85	39	Miss	Miss
1998.11.05	19:00	360	1118	-4.075720734	18	72	55	Hit	Hit
1999.05.03	05:36	360	1584	-4.356547314	-32	56	26	Hit	Miss
1999.05.27	11:36	360	1691	-5.397940009	-78	18	7	Miss	Miss
1999.06.04	06:52	150	2230	-4.408935382	69	19	11	Miss	Hit
2000.02.18	08:38	118	890	-5.853871972	26	20	6	Miss	Miss
2000.04.04	15:12	360	1188	-5.013228274	66	53	29	Miss	Hit
2000.05.15	15:46	165	1212	-5.107905387	67	32	15	Miss	Miss
2000.06.06	14:58	360	1119	-3.638272173	-18	42	27	Hit	Hit
2000.06.10	16:40	360	1108	-4.283996672	38	39	22	Miss	Hit
2000.06.25	07:17	165	1617	-4.721246404	55	64	35	Miss	Miss
2000.07.12	18:41	101	820	-4.244125159	64	26	8	Miss	Miss
2000.07.14	10:03	360	1674	-3.301029996	7	40	21	Hit	Hit
2000.07.22	11:17	259	1230	-4.522878745	56	45	17	Miss	Miss
2000.08.12	13:48	117	499	-5.494850015	46	199	162	Miss	Miss
2000.09.12	11:31	360	1550	-5	9	102	42	Hit	Hit
2000.10.16	06:40	360	1336	-4.602059991	90	151	48	Hit	Hit
2000.10.25	08:45	360	770	-5.397940009	66	396	160	Hit	Hit
2000.11.08	22:42	170	1738	-4.15490196	77	83	46	Hit	Hit
2000.11.24	04:55	360	1289	-3.698970004	5	13	7	Hit	Hit
2000.11.24	14:51	360	1245	-3.638272173	7	30	22	Hit	Hit
2000.11.25	00:59	360	2519	-4.096910013	-50	62	32	Hit	Hit
2001.01.21	19:17	213	664	-5.522878745	-36	22	11	Miss	Miss
2001.01.28	15:40	360	916	-4.823908741	59	44	20	Miss	Miss
2001.03.25	16:25	360	677	-5.045757491	-25	45	11	Miss	Miss
2001.03.29	09:57	360	942	-3.769551066	19	35	18	Miss	Hit

Table A1. Continued

SXR Date	SXR Start Time (hh:mm)	CME width, s (^{o})	CME Velocity, u (km s^{-1})	Logarithm of SXRs $\log SXRs$	Solar Flare position, Lon (^{o})	Solar Flare Duration, DT (min)	Solar Flare Rise time, RT (min)	SEP Forecast Result	SEP Forecast Result
2001.04.02	10:58	50	992	-4	62	67	38	Miss	Miss
2001.04.02	21:32	100	2505	-2.698970004	82	31	19	Hit	Hit
2001.04.09	15:20	360	1192	-4.102372903	4	40	14	Miss	Miss
2001.04.10	05:06	360	2411	-3.638272173	9	36	20	Hit	Hit
2001.04.12	09:39	360	1184	-3.698970004	43	70	49	Hit	Hit
2001.04.15	13:19	167	1199	-2.841637519	85	36	31	Hit	Hit
2001.04.18	02:11	360	2465	-5.65757731	115	5	3	Hit	Hit
2001.04.26	11:26	360	1040	-4.15490196	31	113	106	Hit	Hit
2001.05.20	06:00	179	546	-4.19382002	89	6	3	Miss	Miss
2001.06.15	16:15	360	1701	-5.65757731	-18	11	5	Miss	Miss
2001.09.15	11:04	130	478	-5	49	50	24	Miss	Miss
2001.09.24	09:32	360	2402	-3.585026668	-23	97	64	Hit	Hit
2001.10.01	04:41	360	1405	-4.040958589	91	42	34	Hit	Hit
2001.10.09	10:46	360	973	-5	-8	63	27	Miss	Miss
2001.10.19	00:47	360	558	-4	18	26	18	Miss	Miss
2001.10.19	16:13	360	901	-3.795880011	29	30	17	Miss	Hit
2001.10.22	14:27	360	1336	-4.17392521	-18	64	41	Hit	Hit
2001.11.04	16:03	360	1810	-4	18	54	17	Hit	Hit
2001.11.17	04:49	360	1379	-4.552841976	-42	82	36	Miss	Miss
2001.11.22	20:18	360	1443	-4.420216409	67	34	18	Hit	Hit
2001.12.11	07:58	121	891	-3.552841976	-41	16	10	Miss	Miss
2001.12.26	04:32	212	1446	-4.148741657	54	135	68	Hit	Hit
2001.12.30	13:42	290	718	-5.494850015	43	17	5	Miss	Miss
2002.03.15	22:09	360	957	-4.698970004	3	153	61	Miss	Miss
2002.04.17	07:46	360	1240	-4.585026668	34	131	38	Hit	Hit
2002.04.21	00:43	360	2393	-4	84	115	68	Hit	Hit
2002.07.15	19:59	360	1151	-3.522878745	1	15	9	Hit	Hit
2002.07.20	21:04	360	1941	-3.481486066	-90	50	26	Hit	Miss
2002.08.14	01:47	133	1309	-4.638272173	54	59	25	Miss	Miss
2002.08.16	11:32	360	1585	-4.301029996	-20	95	60	Hit	Hit
2002.08.22	01:47	48	248	-4.267606233	62	18	10	Miss	Miss
2002.08.24	00:49	360	1913	-3.522878745	81	42	23	Hit	Hit
2002.09.05	16:18	360	1748	-5.283996672	-28	77	48	Hit	Miss
2002.11.09	13:08	360	1838	-4.337242177	29	28	15	Hit	Hit
2002.11.24	20:14	360	1077	-5.19382002	37	43	15	Miss	Miss
2002.12.19	21:34	360	1092	-4.568636228	10	43	19	Miss	Miss

Table A1. Continued

SXR Date	SXR Start Time (hh:mm)	CME width, s (^{o})	CME Velocity, u (km s^{-1})	Logarithm of SXRs $\log SXRs$	Solar Flare position, Lon (^{o})	Solar Flare Duration, DT (min)	Solar Flare Rise time, RT (min)	SEP Forecast Result	SEP Forecast Result
2003.05.28	00:17	360	1366	-3.522878745	17	22	10	Hit	Hit
2003.05.31	02:13	360	1835	-4.031517043	65	27	11	Hit	Hit
2003.06.17	22:27	360	1813	-4.22184875	-61	45	28	Hit	Miss
2003.10.26	17:21	171	1537	-3.920818737	38	120	58	Hit	Hit
2003.10.28	09:51	360	2459	-2.764471534	-8	93	79	Hit	Hit
2003.10.29	20:37	360	2029	-3	2	24	12	Hit	Hit
2003.11.02	17:03	360	2598	-3.080921898	56	36	22	Hit	Hit
2003.11.04	19:29	360	2657	-2.552841969	83	37	21	Hit	Hit
2003.11.20	07:35	360	669	-4.01772875	8	18	12	Miss	Miss
2003.11.20	23:42	52	494	-4.301029996	17	16	11	Miss	Miss
2003.12.02	09:40	150	1393	-5.142667515	92	14	8	Miss	Miss
2004.04.11	03:54	314	1645	-5.01772875	46	41	25	Miss	Hit
2004.07.22	07:41	151	700	-5.275724115	-10	27	18	Miss	Miss
2004.07.25	13:37	360	1333	-4.65757731	30	18	12	Miss	Hit
2004.07.31	05:16	197	1192	-5.075720734	89	238	101	Miss	Miss
2004.09.12	00:04	360	1328	-4.397940009	-49	89	52	Hit	Miss
2004.10.30	06:08	360	422	-4.376750729	22	14	10	Miss	Miss
2004.11.07	15:42	360	1759	-3.698970004	17	33	24	Hit	Hit
2004.11.10	01:59	360	3387	-3.602059991	49	21	14	Hit	Hit
2004.12.02	23:44	360	1216	-5	2	51	22	Miss	Miss
2005.01.15	05:54	360	2049	-4.065501529	-4	83	44	Hit	Hit
2005.01.15	22:25	360	2861	-3.698970004	5	66	37	Hit	Hit
2005.01.17	06:59	360	2094	-3.420216409	25	188	173	Hit	Hit
2005.01.20	06:36	360	3256	-3.148741657	61	50	25	Hit	Hit
2005.05.06	03:05	109	1120	-5.031517043	71	16	9	Miss	Miss
2005.05.13	16:13	360	1689	-4.096910013	-11	75	44	Hit	Hit
2005.07.13	14:01	360	1423	-4.301029996	90	97	48	Hit	Hit
2005.07.14	10:16	360	2115	-4	98	73	39	Hit	Hit
2005.08.22	00:44	360	1194	-4.698970004	54	94	49	Miss	Hit
2005.08.22	16:46	360	2378	-4.25181198	65	76	41	Hit	Hit
2005.09.13	19:19	360	1866	-3.823908741	-10	98	8	Hit	Hit
2006.07.06	08:13	360	911	-4.602059991	34	38	23	Miss	Miss
2006.12.13	02:14	360	1774	-3.468521071	23	43	26	Hit	Hit
2006.12.14	21:07	360	1042	-3.823908741	46	79	68	Hit	Hit
2010.08.14	09:38	360	1205	-5.356547314	50	53	27	Miss	Miss
2011.03.07	19:43	360	2125	-4.43179827	59	75	29	Hit	Hit

Table A1. Continued

SXR Date	SXR Start Time (hh:mm)	CME width, s (^{o})	CME Velocity, u (km s^{-1})	Logarithm of SXRs $\log SXRs$	Solar Flare position, Lon (^{o})	Solar Flare Duration, DT (min)	Solar Flare Rise time, RT (min)	SEP Forecast Result	SEP Forecast Result
2011.06.07	06:16	360	1255	-4.602059991	54	43	25	Miss	Hit
2011.08.04	03:41	123	338	-4.031517043	36	23	16	Miss	Miss
2011.08.08	18:00	237	1343	-4.455931956	61	18	10	Miss	Miss
2011.08.09	07:48	360	1610	-3.161150903	69	20	17	Hit	Hit
2011.10.22	10:00	360	1005	-5	77	189	70	Miss	Hit
2011.12.25	18:11	125	366	-4.397940009	31	9	5	Miss	Miss
2012.01.19	13:44	360	1120	-4.494850015	-22	246	141	Hit	Hit
2012.01.23	03:38	360	2175	-4.060480757	25	56	21	Hit	Hit
2012.01.27	17:37	360	2508	-3.769551066	71	79	60	Hit	Hit
2012.03.04	10:29	360	1306	-4.698970004	-61	107	23	Miss	Miss
2012.03.07	00:02	360	2864	-3.301029996	-27	38	22	Hit	Hit
2012.03.13	17:12	360	1884	-4.102372903	59	73	29	Hit	Hit
2012.05.17	01:25	360	1582	-4.292429832	76	49	22	Hit	Hit
2012.06.14	12:52	360	987	-5	-5	184	103	Hit	Miss
2012.07.06	23:01	360	1828	-4	51	13	7	Hit	Hit
2012.07.12	15:37	360	885	-3.853871972	1	113	72	Hit	Hit
2012.07.17	12:03	176	958	-5	65	301	192	Hit	Hit
2012.08.31	19:45	360	1442	-5.075720734	-42	126	58	Hit	Miss
2012.09.27	23:36	360	947	-5.522878745	36	58	21	Miss	Miss
2013.01.16	18:21	250	648	-5.65757731	76	104	62	Miss	Miss
2013.03.15	05:46	360	1063	-4.958607305	-12	169	72	Miss	Miss