Open issues in statistical forecasting of solar proton events: a Machine Learning perspective

Mirko Stumpo[1,2], Simone Benella[1], Monica Laurenza[1], Tommaso Alberti[1], Giuseppe Consolini[1]and Maria Federica Marcucci[1]

[1]INAF-Istituto di Astrofisica e Planetologia Spaziali, Via del Fosso del Cavaliere, 100, I-00133 Roma, Italy
[2]Dept. of Physics, University of Rome Tor Vergata, I-00133 Roma, Italy

Key Points:

- We reinterpret the ESPERTA model in the framework of machine learning, apply rare events corrections and perform a suitable cross validation
- We obtain a good performance, especially for central and well-connected events
- We find that the FAR depends on the ratio between the SPE associated and not associated flares, which has to be considered in the validation

Corresponding author: Monica Laurenza, monica.laurenza@inaf.it

Abstract

Several techniques have been developed in the last two decades to forecast the occurrence of Solar Proton Events (SPEs), mainly based on the statistical association between the >10 MeV proton flux and precursor parameters. The Empirical model for Solar Proton Events Real Time Alert (ESPERTA, Laurenza et al., 2009) provides a quite good and timely prediction of SPEs after the occurrence of ≥M2 X-ray bursts, by using as input parameters the flare heliolongitude, the soft X-ray and the ∼1 MHz radio fluence. Here, we reinterpret the ESPERTA model in the framework of machine learning and perform a cross validation, leading to a comparable performance. Moreover, we find that, by applying a cut-off on the ≥M2 flares heliolongitude, the False Alarm Rate (FAR) is reduced. The cut-off is set to E20 where the cumulative distribution of ≥M2 flares associated with SPEs shows a break which reflects the poor magnetic connection between the Earth and eastern hemisphere flares. The best performance is obtained by using the SMOTE algorithm, leading to probability of detection of 0.83 and a FAR of 0.39. Nevertheless, we demonstrate that a relevant FAR on the predictions is a natural consequence of the sample base rates. From a Bayesian point of view, we find that the FAR explicitly contains the prior knowledge about the class distributions. This is a critical issue of any statistical approach, which requires to perform the model validation by preserving the class distributions within the training and test datasets.

1 Introduction

Solar proton events (SPEs) constitute a major Space Weather hazard in the interplanetary and near-Earth space, as they can hamper spacecraft operations, damage satellites instruments and disrupt radio communications in the Earth atmosphere, as well as pose a radiation threat for astronauts and crews and passengers of airlines in polar routes. Hence, a warning system is required in order to predict SPEs occurrence and mitigate their effects.

Several empirical SPE forecasting models have been developed, which are mainly based on statistical association between the > 10 MeV proton flux and precursor solar parameters or measurements of fast-arriving particles at 1 AU. The first quasi-operational SPE forecasting technique was the proton prediction system (PPS76, developed and improved by Smart and Shea (1979, 1989) and validated by Kahler et al. (2007), which is driven by solar flare parameters (either microwave or X-ray) and flare location, and gives as output an SPE time-intensity profile. Another long-standing model is Protons, currently in use at NOAA SWPC (although including a forecaster in the loop), that uses the time-integrated soft X-ray flux, peak soft X-ray flux, and the location of the associated flare as input parameters, and additionally, the occurrence or non-occurrence of metric radio type II and type IV bursts, indicating the presence of a CME driven shock. The Protons model predicts the probability of a ≥ 10 MeV proton event, the delay time until onset, and the time of the maximum, all with respect to the maximum of the X-ray flare. Other models rely on CME and shock related parameters, such as Winter and Ledbetter (2015) or FORSPEF (Papaioannou et al., 2016), St. Cyr et al. (2017).

Other forecasting schemes such are REleASE and UMASEP esploit, respectively, the early arrival at 1AU of relativistic electrons and protons with respect to lower energy protons at the Earth. Originally, REleASE relied on realtime data from the Solar and Heliospheric Observatory, and more recently has been adapted for use with the Advanced Composition Explorer (Malandraki & Crosby, 2018; Núñez, 2018). UMASEP is based on the correlation between the first derivative of the soft X-ray flux and the first derivative of at least one of the GOES differential proton flux channels. The algorithm looks for the onset of high-energy particles, after an X-ray burst and it makes a prediction about the subsequent evolution of the event based on an empirical relationship between the GOES X-ray flux and the GOES energetic proton flux channels.

The Empirical model for Solar Proton Events Real Time Alert (ESPERTA) model (Laurenza et al., 2009; Alberti et al., 2017) is based on the logistic regression analysis on three solar parameters, viz., the flare location, 1-8 Å soft X-rays (SXR) and 1 MHz Type III fluences (SXR fluences are a measure of flare size/energy and Type III solar radio bursts are the signatures of fast electron beams streaming outward, along the open or quasi open field lines). A prime focus of the ESPERTA model was to provide a timely warning within 10 minutes following the SXR peak for $\geq$ M2 flares. Moreover, the ESPERTA model had been adjusted (Laurenza et al., 2018) to provide early forecasts of the largest radiation storms which are produced by $\geq$ 100 pfu SPE events, once the >10 MeV proton flux crosses the 10 pfu threshold at the Earth. It is worthwhile to note that this is the only model tested to predict $\geq$ 100 pfu (from moderate to extreme) SPE events (with a median warning time of about 2hr) over an extended dataset covering the period 1995-2014.

In recent years, the machine learning (ML) approach has become popular in finding patterns in several scientific contexts (Butler et al., 2018; Carleo et al., 2019; Camporeale et al., 2018). In this framework, few studies have been attempted for SPE forecasting. For instance, a decision tree (DT) model was proposed by Boubrahimi et al. (2017) to predict > 100 MeV SPEs by using the GOES soft X-ray (SXR) and high energy proton observations. More recently, Núñez and Paul-Pena (2020) applied the DT model to two of the ESPERTA parameters, i.e., SXR and radio fluence, and claimed a comparable performance with respect to ESPERTA. ML could represent a powerful way to improve our forecasting capabilities, but the potential application to SPE prediction needs to be assessed. On the other hand, it should be kept in mind that correlations found by ML model should not be confused with causation between input and output variables. ML models try to learn the past instead of uncovering the real/causal relationships between variables that will hold over time. As far as SPEs are concerned, a major limitation is represented by the paucity of SPE associated flares with respect to the not associated ones, being their ratio less than 20% (Laurenza et al., 2007). In ML, prediction of rare events is closely related to the problem of imbalancing, which in principle can be overcome for instance through oversampling techniques to balance properly the ratio of class cardinalities to a fixed value.

In this paper we perform a machine learning approach to SPE forecasting and reinterpret the ESPERTA model, by using the logistic regression. We also apply rare-events corrections to possibly address the problem of the SPE associated flare dataset imbalancing with respect to non SPE associated flares one. We also perform a suitable cross validation and find the conditions to obtain the best performance of the method, i.e. for intermediate and well connected flare longitudes (greater than E20°).

Section 2 presents the basics of the ML approach as well as a description of the logistic regression technique and cross validation method. In Section 3 we perform the application to SPE forecasting and validate the method. We also discuss the effects of imbalancing by interpreting the natural distribution of the SPE events from a Bayesian point of view. Finally, section 4 discusses the comparison with competing techniques and draws the conclusions.

2 The Supervised Learning Approach

In the framework of supervised learning paradigm, the problem to be solved is: given a series of examples of the target variable t associated to a set of certain input variables (features) X, we want to train a model $\mathcal{M}$ which is able to predict the value of t for any out-of-sample data point. Roughly speaking, suppose to have an object described by its n_f features, and suppose to have its status (target, e.g. 1 or 0) such that each set of features are associated to a given status of the object. The objective of ML approach is to find a pattern, into the n_f dimensional space of the features, which characterize the sta-

tus of the object. The target variable t can be either continuous or categorical, being respectively the cases of *regression* and *classification* problems. In the latter case, the possible status assumed by the target variable are called *classes*. Since SPE forecasting is recast in a classification problem, from now on we will cover only this case study.

In order to fit a model, the ML algorithm takes, as input, a model whose weights $\mathbf{w} = (w_1, w_2, ..., w_{n_f})$ are optimized with respect to the training set, so that learning the optimal set of weights $\mathbf{w}^{(opt)}$ to be put into the final model. Roughly speaking, the weight w_i can be associated to the importance of the i-th feature. From a mathematical point of view, ML is an n_f-dimensional optimization problem and the model $\mathcal{M}$ can be either a function $f(\mathbf{w}, \mathbf{x})$ mapping the feature vector $\mathbf{x} = (x_1, x_2, ..., x_{n_f})$ in its class object $t \in \mathbb{N} = \{0, 1, 2, \ldots, N\}$ or a probability distribution function $P(\mathcal{C}_i|\mathbf{x}, \mathbf{w})$ representing the probability of observing the i-th class $\mathcal{C}_i$ given the feature and weight vectors. The optimization problem is solved with respect to an error function $L(f, t)$ (or $L(P, t)$ if using probabilistic model), called loss function, measuring the distance between the estimations given by the model f and the optimal target t ((Friedman et al., 2001)), i.e.

$$f^{(opt)} = \arg\min_{f \in F} L(f, t), \tag{1}$$

where F is a function space instead of a parameter space. But, if the model is fixed, i.e. for instance $f = f(\mathbf{x}, \mathbf{w})$, the problem is recast in parameters optimization which is much easier to solve:

$$f^{(opt)} \rightarrow \mathbf{w}^{(opt)} = \arg\min_{\mathbf{w} \in D_w} L(f_{\mathbf{w}}, \mathbf{x}, t), \tag{2}$$

where now D_w is the space of parameters. There are two different approaches to the classification problem (Bishop, 2006). The simplest involves the construction of a discriminant function that directly assigns each input vector $\mathbf{x}_i$ to a specific class. A more powerful approach models the conditional probability distribution $P(\mathcal{C}_k|\mathbf{x}_i, \mathbf{w})$, i.e. the probability of observing the k-th class given the input vector and the weight vector. In the first case our model is the discriminant function $f(\mathbf{x}_i, \mathbf{w})$ mapping the input vector in its class, while in probabilistic classification the model is a probability function $P(\mathcal{C}_k|\mathbf{x}_i, \mathbf{w})$. Both deterministic and probabilistic approaches are associated to a decision rule; whereas for deterministic approach the decision rule is given by f itself, for the probabilistic approach we associate a probability threshold ϵ which, in the simple case of binary classification, maps the target variable as follow:

$$t(\mathbf{x}_i) = \begin{cases} 1 & \text{if } P(\mathcal{C}_1|\mathbf{x}_i, \mathbf{w}) \geq \epsilon \\ 0 & \text{if } P(\mathcal{C}_1|\mathbf{x}_i, \mathbf{w}) < \epsilon \end{cases}. \tag{3}$$

The part in which we fit the model by solving eqs. 1 or 2 is called training phase. After having learned the model, it is tested on out-of-sample dataset; this part is called test phase. In literature can be found a lot of algorithms (e.g. random forest (Breiman, 2001), decision tree (Quinlan, 1996), logistic regression (Cox, 1958; Cramer, 2002), k-th nearest neighbour classificator (Fix, 1985), support vector machines (Boser et al., 1992), etc.) which implement both their own learning and prediction rules. Generally, the probabilistic approach is more powerful in respect to the direct classification algorithms because it allows to interpret the risk associated to our decision. The decision function of the machine learning algorithms can be recast into probabilities but they are known to not be well calibrated (Collett, 2002; Niculescu-Mizil & Caruana, 2005). On the other hand, the estimation given by the logistic regression is a true probability mapping (Lichtenstein et al., 1977), being not just pseudo-probability as produced by many machine learning algorithms. This fact, together with the mathematical properties of the logistic function, represents the major advantage of probabilistic modeling with logistic regression (Maalouf, 2011; King & Zeng, 2001), thus in this work we will use this approach, which will be discussed in the next section.

2.1 The Logistic Regression

In order to fit the best model, the logistic regression algorithm uses the discriminative approach (Bishop, 2006; Hosmer Jr et al., 2013), i.e. we assume that

$$P(\mathcal{C}_1|\mathbf{x},\mathbf{w}) = \frac{1}{1+e^{-(w_0+\mathbf{w}^T\mathbf{x})}} \tag{4}$$

from which we obtain directly

$$P(\mathcal{C}_0|\mathbf{x},\mathbf{w}) = 1 - P(\mathcal{C}_1|\mathbf{x},\mathbf{w}) = \frac{1}{1+e^{w_0+\mathbf{w}^T\mathbf{x}}}, \tag{5}$$

where w_0 is a constant of the model. From now on, we set $\mathbf{w} = (w_0, w_1, ..., w_{n_f})$ and $\mathbf{x} = (1, x_1, ..., x_{n_f})$. In order to optimize the weights $\mathbf{w}$, we use the maximum likelihood estimation (MLE) (McCullagh & Nelder, 1989). The likelihood function $L(\mathbf{X}|P)$ is essentially the probability associated with the observed dataset once the model $P = P(\mathcal{C}_1|\mathbf{x},\mathbf{w})$ is given. Thus, maximizing the likelihood function corresponds to finding those weights maximizing the probability of the observed dataset. It is expressed as the product of the probabilities of N individual observations (Bernoulli trials), where N is the total number of training data points i.e.

$$L(\mathbf{X}|P) = \prod_{t_i=1} P(\mathcal{C}_1|\mathbf{x}_i,\mathbf{w}) \prod_{t_i=0} (1 - P(\mathcal{C}_1|\mathbf{x}_i,\mathbf{w})), \tag{6}$$

where: 1) $t_i \in \{0,1\}$ is a variable such that $t_i = 1$ if $\mathbf{x}_i \in \mathcal{C}_1$ and $t_i = 0$ if $\mathbf{x}_i \in \mathcal{C}_0$; 2) $\mathbf{X}$ is the whole (training) dataset, i.e. a $(N \times n_f)$ matrix whose rows correspond to each observation vector $\mathbf{x}_i$. Note that the first column of $\mathbf{X}$ is a vector of ones. Since the logarithm is a monotonic function preserving the position of maxima, in order to handle the products we prefer to maximize the log-likelihood function, which gives the negative cross-entropy error function

$$\mathcal{L}(\mathbf{X}|P) = \sum_{t_i=1} \log P_i + \sum_{t_i=0} \log(1 - P_i). \tag{7}$$

Now, the gradient with respect to the weights $\mathbf{w}$, by keeping constant $\mathbf{x}_i$, results in the system of equations

$$\nabla\mathcal{L}(\mathbf{X}|P) = \sum_{i=1}^{N} [t_i(1-P_i) - (1-t_i)P_i]\,\mathbf{x}_i = \sum_{i=1}^{N} (t_i - P_i)\,\mathbf{x}_i = 0, \tag{8}$$

where we kept together summation for $t_i = 0$ and $t_i = 1$. The last equation can be rewritten in matrix form, i.e.

$$\nabla\mathcal{L}(\mathbf{X}|P) = \mathbf{X}^T(\mathbf{t} - \mathbf{P}) = 0, \tag{9}$$

where $\mathbf{P}$ is the probability vector associated to each observation.

Thus, the optimization problem is recast in solving a system of n_f equations with respect to the weights $\mathbf{w}$. For logistic regression there is not a closed-form solution due to non-linearity of the logistic sigmoid function but this is not a real problem since it can be proved that the negative cross-entropy is a convex function having a unique maximum (Bishop, 2006). Furthermore, the solution can be found by means of the Newton-Raphson method, yielding the following iterative set of equations

$$\mathbf{w}^{(new)} = \mathbf{w}^{(old)} + \mathcal{H}^{-1}\nabla\mathcal{L}(\mathbf{X}|P), \tag{10}$$

where $\mathcal{H}$ is the Hessian matrix of $\mathcal{L}$ whose elements are the second derivatives of Equation (7) with respect to $\mathbf{w}$. Performing calculations it can be proved that the iterative equations can be rewritten as

$$\mathbf{w}^{(new)} = \mathbf{w}^{(old)} - (\mathbf{X}^T\mathbf{R}\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{P} - \mathbf{t}), \tag{11}$$

where $\mathbf{R}$ is the diagonal matrix such that $R_{ii} = P_i(1 - P_i)$. Henceforth, we will call the method yielded by eq 11 the *basic MLE*.

2.2 Rare events and imbalanced datasets

In the framework of rare events, the class frequencies are imbalanced, i.e. $n_0/N = p(\mathcal{C}_0) \gg p(\mathcal{C}_1) = n_1/N$, leading their classification to be quite challenging. In this particular case study the limitation has its root cause in the process generating the event itself, causing the predictive model to be biased toward the majority class (King & Zeng, 2001; Gao & Shen, 2007). The problem is amplified when we have a small dataset; indeed it is known that the MLE method, used for weights optimization in the previous section, is an asymptotically consistent estimator, i.e. it is unbiased only when applied to large datasets (Collett, 2002).

In this work we applied and compared the following methods in order to reduce the bias induced by the imbalancing:

1. *Weighted MLE*: The error induced by the misclassification of minority class is increased operating directly on the loss function (Manski & Lerman, 1977). In other words, if the original loss function is $\mathcal{L} = \mathcal{L}_0 + \mathcal{L}_1$, where $\mathcal{L}_0$ and $\mathcal{L}_1$ are the loss functions associated respectively to the class $\mathcal{C}_0$ and $\mathcal{C}_1$, then the class weighted loss will be $\mathcal{L} = K_0 \cdot \mathcal{L}_0 + K_1 \cdot \mathcal{L}_1$, so that the Equation (7) is rewritten as

$$\mathcal{L}_K(\mathbf{X}|P) = K_0 \cdot \left(\sum_{i=1,t_i=1}^{N} \log P_i \right) + K_1 \cdot \left(\sum_{i=1,t_i=0}^{N} \log\left(1 - P_i\right) \right) \tag{12}$$

   and the Newton-Raphson algorithm is applied to the gradient function

$$\nabla \mathcal{L}_K(\mathbf{X}|P) = \sum_{i=1}^{N} \left[(f \cdot t_i - P_i) + t_i P_i (1 - f) \right] \mathbf{x}_i, \tag{13}$$

   where $f = K_1/K_0$. Note that if $f = 1$, this equation yields Equation (9). In order to balance the classes, the weights C_i are chosen to be inversely proportional to the class frequencies in the dataset. By applying this correction to the loss function the errors induced by the misclassification of both classes are balanced;
2. *SMOTE MLE*: We create a synthetic set of events with the Synthetic Minority Oversampling Technique (SMOTE), which works by selecting two nearby points in the feature space and generating a new point between them (Chawla et al., 2002). Let $\mathbf{x}_i \in \mathcal{C}_1$ and $\mathbf{x}_j \in \mathcal{C}_1$ be two feature vectors, then the synthetic data vector $\mathbf{x}_s$ is given by

$$\mathbf{x}_s = \mathbf{x}_i + \lambda \left(\mathbf{x}_j - \mathbf{x}_i \right), \tag{14}$$

   where λ is a random number sampled from a uniform distribution $\mathcal{U}(0, 1)$. This method is an improvement of the Random Oversampling Technique which simply creates duplicated copies of original data points. We emphasize that the oversampling technique must be used only over the training set in order to avoid optimistic evaluations. Indeed, the synthetic points are similar to the original ones so that resulting too simple to predict. It means that, before applying oversampling, the original dataset must be split in training and test datasets. Then a fixed number of minority class data $n_1^{(train)}$ is randomly put into the training set and used to generate synthetic data, while the remaining part is conserved into the test set. Thus all the synthetic data are kept into the training dataset and not used for testing. Once the training dataset has been filled with synthetic samples, the model is obtained, in the particular case of logistic regression, by means of basic MLE in eq 11.

2.3 Testing the model

Once the optimal weights have been learned from data solving the eqs. 1 or 2, we have to test the ability of the model to make predictions. Clearly, it cannot be done with respect to the same sample used for the training phase because it will results in a too optimistic evaluation. Thus the basic idea in applications is to split up the original dataset into a training and testing datasets; then the prediction power of the model is referred to its ability to generalize what it has learned during the training phase; the error induced by the inability of the model to generalize the prediction power is known as *overfitting* (Dietterich, 1995). In this framework, the goodness of the model is always a trade-off between the optimization of the loss function and the optimization of the scores computed during the test phase.

The ability of the model to make predictions is measured by the confusion matrix

$$C = \begin{pmatrix} \mathrm{TN} & \mathrm{FN} \\ \mathrm{FP} & \mathrm{TP} \end{pmatrix}, \tag{15}$$

whose elements are: the total number of true negatives (TN, i.e. the correct nulls), true positives (TP, i.e. the hits), false negatives (FN, i.e. the misses) and false positives (FP, i.e. the false alarms) computed with respect to the test set. From these numbers we compute the Probability Of Detection (POD) and the False Alarm Rate (FAR), which are defined as

$$\mathrm{POD} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}, \tag{16}$$

$$\mathrm{FAR} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TP}}. \tag{17}$$

In our application, POD and FAR must be optimized simultaneously, thus we will refer the decision rule with respect to the F_1 and the Critical Success Index (CSI) defined respectively as

$$F_1 = 2 \cdot \frac{\mathrm{POD} \cdot (1 - \mathrm{FAR})}{\mathrm{POD} + \mathrm{FAR}}, \tag{18}$$

$$\mathrm{CSI} = \frac{\mathrm{POD} \cdot (1 - \mathrm{FAR})}{1 - \mathrm{FAR} \cdot (1 + \mathrm{POD})}. \tag{19}$$

Since the performance of the model can depend upon the particular portion of the dataset used for training and testing (e.g. we could have selected, as a test set, a portion too simple to predict), especially when the dataset is poor, then the process of training and testing phases must be re-iterated in order to have a better idea of the performances on different portions of the dataset. The basic approach to get an unbiased estimation of the model performance is the so-called k-fold cross validation (Browne, 2000; Stone, 1978). In this framework, training and test phases are re-iterated k times, so that the model is tested on different portions of the dataset for each iteration, being the test sets non-overlapping and independent for each fold. Each fold is selected randomly from the original dataset.

However, this approach is critical when the classes are not balanced. Indeed, in this case, as we will show later, it is extremely important to preserve the original class distributions within the test and training datasets in each fold.

3 Application to SPE forecasting

In order to set up the ML based scheme for SPE prediction, we first selected a proper feature space to be associated with our target variable, which is the occurence of SPE. Thus, our target variable t is naturally binary, having status 1 if a SPE occur, whereas status 0 if it does not. The ESPERTA forecasting model, designed to work in real time, rely on the $\geq M2$ SXR bursts and type III radio emission, which have proven to be good

indicators of an impending SPE (Laurenza et al., 2009; Alberti et al., 2017), as they are proxies of the flare importance and duration of the particle escape, respectively. In more detail, the time integrated SXR and 1 MHz flux were computed in such a way to be timely available 10 minutes after the flare peak time, in order to maximize the warning time. Moreover, three different longitude ranges for the flare location were considered to take into account the particle propagation from the solar source. Hence, we assumed the three ESPERTA inputs as the set of features ($n_f = 3$) associated to the SPE occurrence.

3.1 Description of the dataset

In this study we started from the list of $\geq$ S1 SPEs already published in (Laurenza et al., 2009) for the period 1995-2005 and in (Alberti et al., 2017) for the period 2006-2014. We updated the dataset by considering a few corrections to the latter as reported in (Laurenza et al., 2018) and then by extending the event list until June 2017 to cover almost entirely the solar cycle 24. The 5 minute averaged proton flux gathered on board the Geostationary Operational Environmental Satellite (GOES) spacecraft series is used in the classification of SPE events. The requirement for the identification of a $\geq$ S1 event is to observe a ≥ 10 MeV ≥ 10 pfu flux for at least three data points. In compiling the dataset we took into account the NOAA SPE event list although it was corrected for several events. In particular, we included as separate events subsequent SPEs when the intensity lies above 10 pfu.

We obtained 92 SPE associated with $\geq$ M2 flares, 21 SPE associated with $<$ M2 flares over the period January 1995 - April 2017. All SPE events are listed in Appendix A.

In order to apply the ML approach to the SPE forecasting, we also considered all the 989 $\geq$ M2 flares occurred during the same time period, as they have been shown to be well associated with the occurrence of a SPE. In particular, we described each $\geq$ M2 by the three features ($n_f = 3$) used in the ESPERTA model.

The SXR time integrated flux (I) is calculated from the 1/3 power point before the peak to the 1/3 power point after. If the X-ray intensity drops by a factor 3 within 10 minutes of the peak, the integration stops, otherwise an exponential fit of the flare is used to extrapolate the intensity curve to the 1/3 power point. The fit is based on the intensity values from 6 to 10 minutes after the peak and it is a reasonable tool to take into account the flare profile. The radio time integrated flux (J) is computed by integrating the 1MHz flux from 20 minutes before the time of the 1/3 X-ray peak until 10 minutes after the X-ray peak.

Out of the 989 flares, 933 do not contain data gaps. The final dataset contains a total of $n_1^{(total)} = 92$ SPE-associated flares herafter referred to as SPE class and 842 non-associated flares herafter referred to as NO SPE class, corresponding to an imbalancing between the two classes of 1 : 9.

Normalized probability densities of the three features are displayed in Figure 1 for both SPE and NO SPE classes. Some differences can be highlighted, meaning that our input variables are correlated to the occurrence of SPEs:

1. SPE events privilege western longitude (Figure 1, left panel) of the observed flare, while NO-SPE events are uniformly distributed. As we will show later, this behaviour is due to the fact that eastern SPE events are not magnetically well connected to the Earth;
2. X-rays Fluence (Figure 1, central panel) for SPE events shows a shift toward greater values meaning that in general a SPE event is associated to greater values of the X-rays Fluence. Nevertheless, the probability of observing a NO-SPE events with an high values of the X-rays fluence is not negligible;

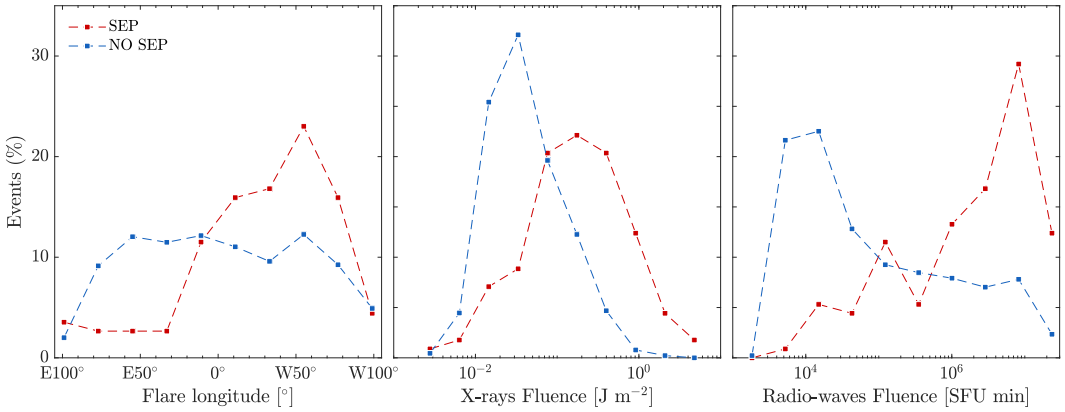


Figure 1. Density plot of the input variables separated by SPE-associated (blue) and non-SPE-associated events (red).

3. Radio-waves Fluence (Figure 1, right panel) shows similar behaviour as that of X-rays Fluence. In particular, the distribution of NO-SPE events, displays an high tail toward higher values comparable to those of SPE events.

In general, the misclassification rates of the events depend upon the overlap between the distribution, being the events lying in the tail of the distributions much more difficult to predict. This effect is amplified by the imbalancing: a false positive in this case affects greatly the goodness of our predictions.

The eastern region of the Sun (negative longitude) produces few SPE-associated events. This fact can be highlighted by computing the cumulative distribution function of the longitude values L for SPE and non-SPE events, defined by

$$P(\tilde{L} < L) = \int_{-\infty}^{L} p(\tilde{L}) d\tilde{L}, \tag{20}$$

where the variable L maps the longitude of the associated flare. In this case, $L < 0°$ and $L > 0°$ stands for the eastern and western region of the Sun respectively. As can be seen in Figure 2, the probability of observing a SPE in a flare with longitude $< L$ (left panel) is quasi-flatten for $L < -20°$, while the distribution of non-SPE events (right panel) is essentially uniform. In particular, the distribution associated to SPE events has a transition at $L = -20°$ longitude value, where a steeper rise of $P(\tilde{L} < L)$ begins. This effect can be due to the fact that eastern SPE are not labeled correctly (even if there is effectively an acceleration phenomenon) because it is a region not magnetically well connected to the Earth, so that SPE events are less easily detected. It is extremely important to take into account this aspect properly since eastern flares mislead the model producing a lot of false positives (and, as we will show later, the challenge of SPE forecasting is driven mainly by the problem of false positives).

Out of a total of 933 flares in our dataset, 332 have a longitude $L \leq -20°$. For these flares we can compute the anti-cumulative distribution functions associated to SXR fluence and radio time integrated flux, i.e. the probabilities $P(\tilde{I} > I) = 1 - P(\tilde{I} < I)$ and $P(\tilde{J} > J) = 1 - P(\tilde{J} < J)$. In other words, our idea is to compare the (right) tails of the distributions of non-SPE events. As it is shown in Figure 3, the X-rays fluence of NO SPE flares in the region $L \leq -20°$ has an higher tail with respect to that for $L > -20°$. Roughly speaking, $L \leq -20°$ non-SPE events present many points with X-rays fluence comparable to that of SPE associated events. Viceversa, for the radio fluence J we have that the central region of the distribution when $L \leq -20°$ is lower than

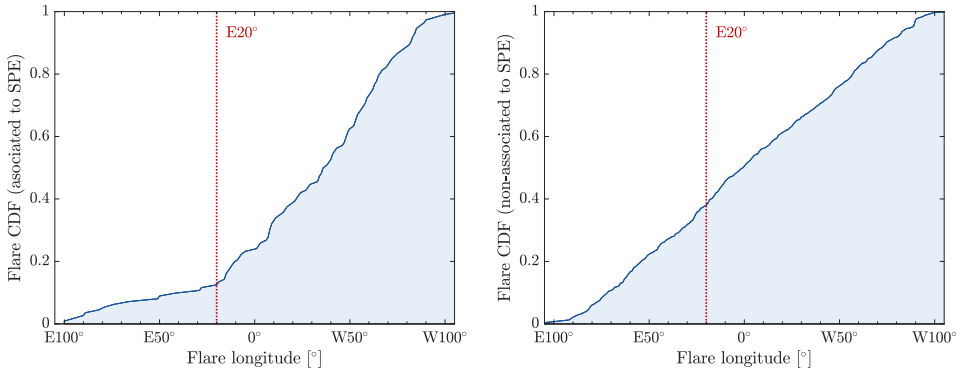


Figure 2. Longitudinal cumulative distribution of SPE-associated flare (left) and non SPE-associated flare (right). Red vertical line selects the longitude for which we have a sharp transition of the probability of observing a SPE event.

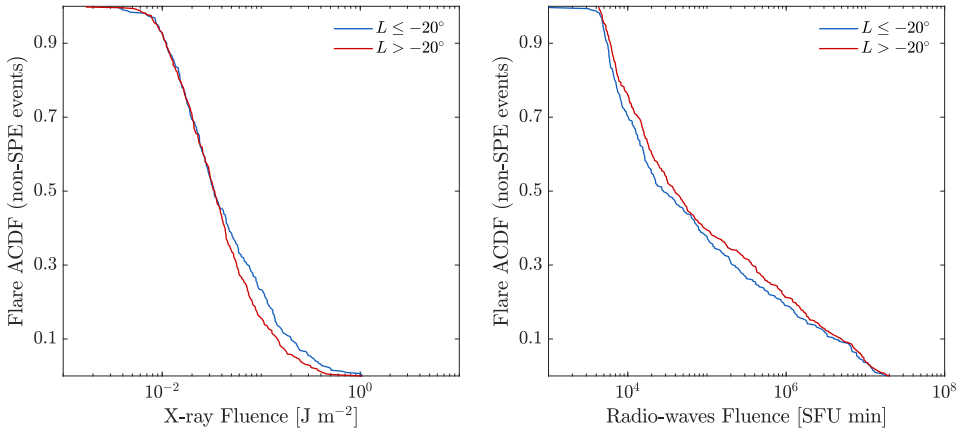


Figure 3. Anti-cumulative distributions of non-SPE events fluxes filtered by longitude.

that corresponding to $L > -20°$, while the tails (i.e. the events which can produce false positives) are essentially equal.

In order to deepen the differences induced by the evidences described above, we will compare the forecasting performances of our model when either the filter to $L > -20°$ is applied or not. In the first case we end up with a total of 601 events corresponding to an imbalancing of $1.3 : 8.7$; in the second case we end up with a total of 933 events and an imbalancing of $1 : 9$.

3.2 The effects of imbalancing

In order to show the effects of class distributions on inducing biased evaluations, firstly we create training and test datasets such that: 1) the fraction τ between SPE and non-SPE events in the test set spawns from 0.1 to 0.5 and 2) the fraction of training samples with respect the total spawns from 0.7 to 0.9. Then the POD and FAR scores have been averaged (over 10 iteration) for each pair of τ and $\tilde{N}$. For these purposes, we kept the decision threshold ϵ in Equation (3) to be equal to 0.5, we used the weighted MLE and the filtering in longitude is not applied.

Let $n_1^{(total)}$ and $n_1^{(train)}$ be the total number of SPE events and those kept into the training dataset respectively and let N be the number of samples in the training dataset.

By means of the analysis described above we get, for POD and FAR scores, the results in Figure 4. In particular, the Probability of Detection (right panel) shows only random fluctuations with respect to N as expected, whereas it becomes larger as $n_1^{(train)}$ grows, meaning that the model distinguishes better a SPE event when more examples are given. The False Alarm Rate has a regular pattern with respect to N and $n_1^{(train)}$, showing that FAR is almost constant along its contours. In particular we note that, if $n_1^{(train)}$ is fixed,

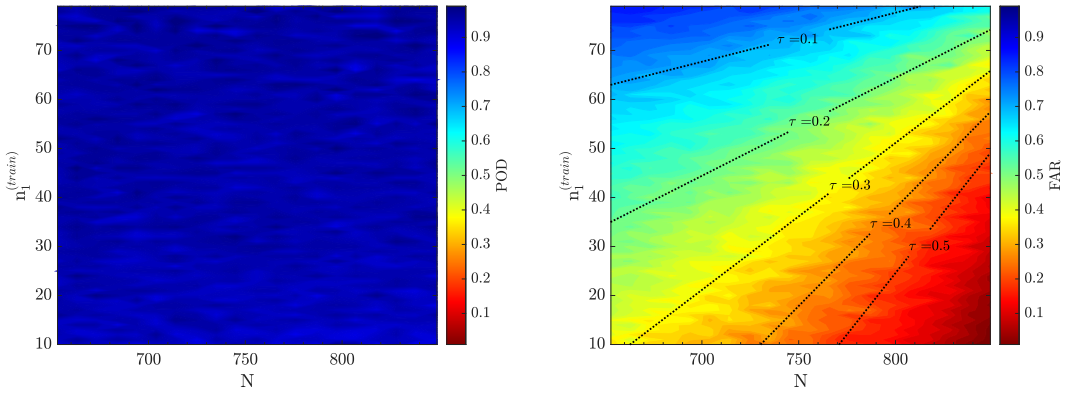


Figure 4. POD (left panel) and FAR (right panel) scores with respect to the number of samples N used for training and the number of SPE events kept into the training dataset. The black dotted lines represent Equation (23) for different values of τ.

the FAR decreases with N. By increasing $n_1^{(train)}$ we note a worsening of the FAR score and the contours become more flatten.

Now we demonstrate that the worsening of the FAR score with respect to these parameters is due to the bias induced by the class distributions. Being $\tilde{N}$ the total number of testing samples, we can define the fraction τ of SPE events in the test dataset as

$$\frac{n_1^{(total)} - n_1^{(train)}}{\tilde{N}} = \tau, \tag{21}$$

from which we get

$$n_1^{(train)} = -\tilde{N}\tau + n_1^{(total)} = -(N_{tot} - N)\tau + n_1^{(total)}, \tag{22}$$

where $N_{tot} = N + \tilde{N}$. Keeping $n_1^{(train)}$ constant while growing N and, viceversa, keeping constant N while lowering $n_1^{(train)}$, means that $\tau \to 0.5$ in the test dataset. Indeed, we found that the contours follow exactly the Equation (22), i.e. the contours are given by

$$n_1^{(train)} = N \cdot \tau + n_1^{(total)} - \tau \cdot N_{tot}. \tag{23}$$

In other words the FAR is constant along the lines which describe a fixed fraction τ between SPE and non-SPE events in the test dataset and it is minimum when $\tau = 0.5$, but clearly this does not correspond to the real situation, since $\tau = 0.5$ corresponds to how the model would perform if there was a probability of 1/2 that a flare had a SPE-associated event. Therefore, in this case τ must always be set equal to the evidence of SPE events, i.e. $\tau = p(\mathcal{C}_{sep})$, which in our dataset correspond to $\tau = 0.15$ when the filter $L > -20°$ is applied and $\tau = 0.1$ when the filter is not applied. Setting $\tau \neq p(\mathcal{C}_{sep})$ will results in biased POD and FAR.

Hence, in our application, it is extremely important to preserve the class distributions within the test and training datasets. Therefore, if the model has been trained on

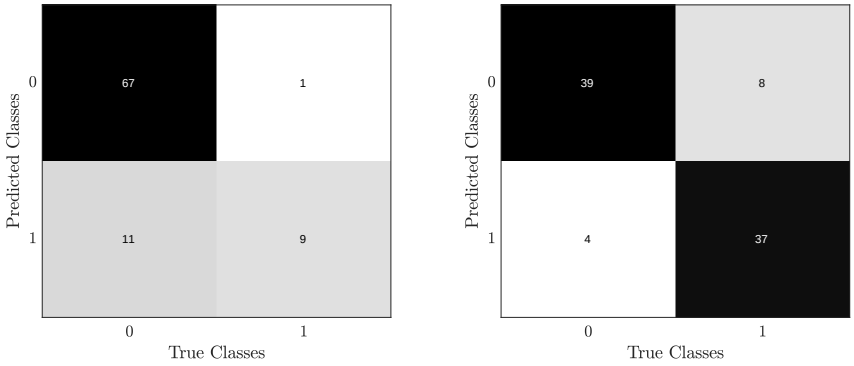


Figure 5. Confusion matrices for unbalanced test (left) and balanced (right) test datasets. In the first case we note that, even if the model predicted 9/10 SPE events (i.e. POD = 90%), and 67/78 NO-SPE events, the FAR score is greater than 0.5. The matrices given in these examples are referred to a particular realization during the cross-validation for two different values of τ.

a balanced dataset, e.g. SMOTE MLE and weighted MLE, the eqs. 16 and 17 must be referred to a test dataset preserving the original class distributions in order to get an unbiased estimation of the model performance. For this reason, in order to cross-validate the model, we constrain each fold to preserve the class distributions between the test and training datasets.

So far we have derived the results in Figure 4 with respect to the logistic regression, but we emphasize that the reasoning which yielded the Equation (22) is also valid for any statistical forecasting tool.

3.3 Model optimization and validation

So far we have considered the default 1/2 threshold in Equation (3), meaning that the model classifies an event as a SPE if $P(\mathcal{C}_{sep}|\mathbf{x},\mathbf{w}) > P(\mathcal{C}_{no-sep}|\mathbf{x},\mathbf{w})$. As we noted previously, there is no reason to keep the threshold ϵ fixed to 1/2 since it defines the decision rule which is separated from the probability inference. Indeed, the estimation given by the logistic regression is true probability mapping, being not just pseudo-probability (or direct classification) as produced by many machine learning algorithms (e.g. decision tree): on the other hand this is the major advantage of probabilistic modeling with logistic regression (Maalouf, 2011; King & Zeng, 2001).

We are interested in minimizing the errors induced by both false positives and false negatives; we reject models with: 1) high POD and high FAR; 2) low POD and low FAR. For instance, the confusion matrix in the left panel of Figure 5 is saying that when a SPE is occurring, it is detected by the model with a probability of 0.9 but, on the other hand, according to the number of false positives, given the model predicted a SPE, there is a probability of 0.55 that is actually a false alarm. This is quite paradoxically but, as we will show later, this can be interpreted from a Bayesian point of view.

In order to choose that ϵ representing the best trade-off between POD and FAR scores as measured by the F_1 and CSI indices in eqs. 18 and 19, we have applied the stratified cross validation method with $k = 10$ folds; the score indices have been averaged over the k folds for each threshold value ϵ. With reference to Figure 6, we found that balancing the class distributions of the training dataset (by weighting or SMOTE) before performing MLE given in Equation (11), corresponds essentially to a translation of the optimal threshold ϵ. In our case of SPE forecasting/classification, the approaches ex-

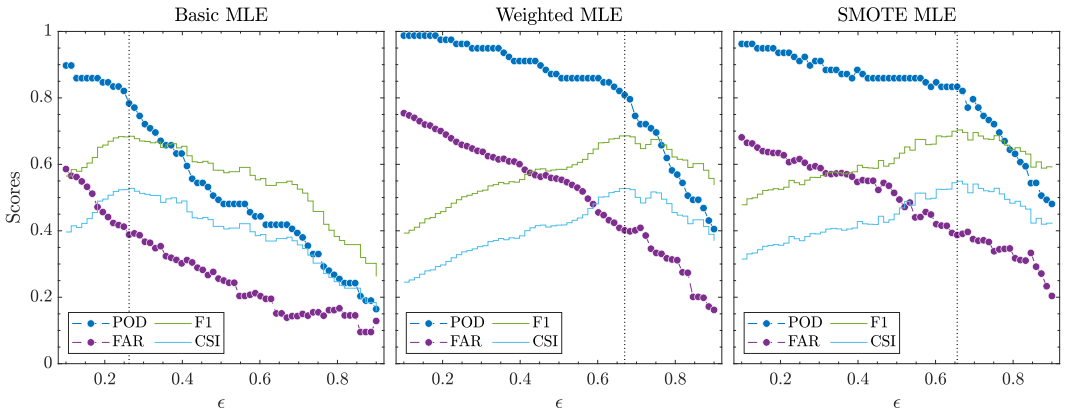


Figure 6. Averaged cross-validated scores with respect to the decision threshold ϵ computed by applying the $L > -20°$ filtering to the flare events. The results are referred to the validation/test set, i.e. the performances on the training set have not been considered.

plore the optimal region, thus there are little differences in terms of prediction power. The scores evaluated at the maximum of the F_1 are reported in Table 2.

By re-introducing the events filtered out, i.e. those events whose flare-associated longitude is $L < -20°$, we are able to test the importance of the hypothesis concerning the longitudinal distribution of SPE events. Using the stratified cross-validation method we get the results shown in Figure 7. Thus, this case highlights a worsening of the pre-

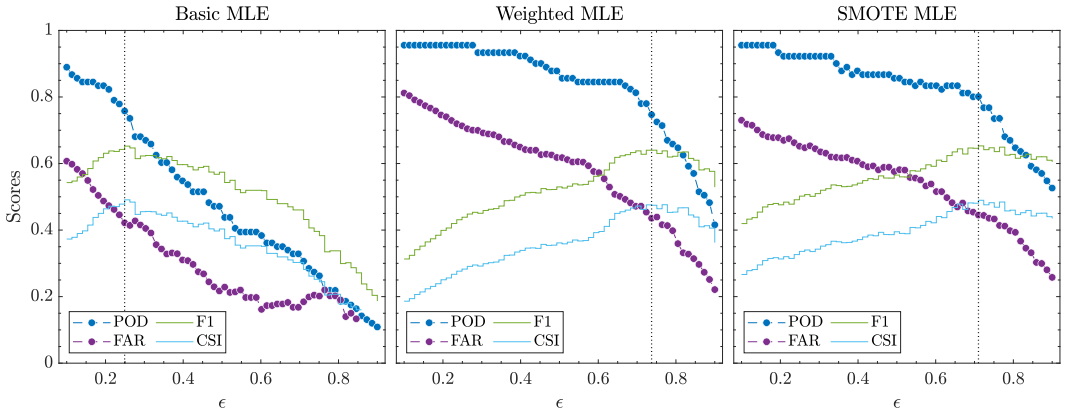


Figure 7. Averaged cross-validated score with respect to the decision threshold ϵ without filtering the events. The results are referred to the validation/test set, i.e. the performances on the training set have not been considered.

diction performances of our model: the optimal results for basic, weighted and SMOTE MLE are reported in Table 2.

By means of those findings and by means of what we have found previously (see figs. 2 and 3) this could be due to the differences in the distributions of non-SPE events for $L \leq -20°$ and $L > -20°$. Finally, it could be an hint that SPE events are labeled wrongly a priori in the region not well magnetically connected to the Earth because they are more difficult to detect.

Table 2. POD and FAR scores of the model by considering the whole dataset and the filtered dataset ($L > -20°$). Model scores are computed for basic, weighted and SMOTE MLE respectively.

		POD	FAR	F1	CSI
Whole dataset	Basic MLE	0.76	0.42	0.65	0.49
	Weighted MLE	0.75	0.44	0.64	0.48
	SMOTE MLE	0.80	0.45	0.65	0.49
$L > -20°$	Basic MLE	0.78	0.39	0.68	0.52
	Weighted MLE	0.81	0.40	0.69	0.53
	SMOTE MLE	0.83	0.39	0.70	0.55

3.4 Bayesian interpretation of the False Alarm Rate

From a Bayesian point of view, we can interpret the POD score as the probability that, given a SPE event, it is labeled as a SPE event by the model, i.e. $\text{POD} = p(\hat{1}|\mathcal{C}_{spe})$. On the other hand, the FAR is the probability that a non-SPE event is labeled as a SPE event, i.e. $\text{FAR} = p(\mathcal{C}_{no-spe}|\hat{1})$, so that the probability of being correct in predicting a SPE event is $1 - \text{FAR} = p(\mathcal{C}_{spe}|\hat{1})$. Whereas the latter equation represents the posterior probability of being correct with respect to the prior knowledge about the SPE distributions, the POD score represents the ability of the model to distinguish a SPE event independently from any prior knowledge.

Using the Bayes' theorem (Grinstead & Snell, 2012) we know that, given the model has predicted a SPE event, the posterior probability that it is actually a SPE event can be written as

$$p(\mathcal{C}_{spe}|\hat{1}) = \frac{p(\hat{1}|\mathcal{C}_{spe}) \cdot p(\mathcal{C}_{spe})}{p(\hat{1})} \tag{24}$$

The overall probability $p(\hat{1})$ of labeling an event as a SPE event can also be obtained from the law of the total probability or, equivalently, from the confusion matrix in 15, i.e.

$$p(\hat{1}) = p(\hat{1}|\mathcal{C}_{spe}) \cdot p(\mathcal{C}_{spe}) + p(\hat{1}|\mathcal{C}_{no-spe}) \cdot p(\mathcal{C}_{no-spe}) = \frac{\text{FP} + \text{TP}}{\text{FP} + \text{TP} + \text{FN} + \text{TN}}, \tag{25}$$

where the probability $p(\hat{0}|\mathcal{C}_{no-spe}) = 1 - p(\hat{1}|\mathcal{C}_{no-spe})$ represents the ability of the model to distinguish a non-SPE event, i.e.

$$p(\hat{1}|\mathcal{C}_{no-spe}) = \frac{\text{FP}}{\text{FP} + \text{TN}} \tag{26}$$

and

$$p(\hat{0}|\mathcal{C}_{no-spe}) = \frac{\text{TN}}{\text{FP} + \text{TN}}. \tag{27}$$

The probabilities $p(\mathcal{C}_i)$ are our prior knowledge of SPE and non-SPE distributions, i.e. their fraction in our sample (test) data being, respectively, τ and $1-\tau$. Hence, the Equation (24) allows to interpret the FAR as a posterior probability with respect the POD and the ratio τ:

$$\text{FAR} = 1 - \frac{\text{POD} \cdot \tau}{\text{POD} \cdot \tau + \frac{\text{FP}}{\text{FP+TN}} \cdot (1 - \tau)}. \tag{28}$$

or, equivalently,

$$\text{FAR} = 1 - \text{POD} \cdot \frac{\tau}{p(\hat{1})}, \tag{29}$$

showing that the FAR score 1) depends explicitly upon τ; 2) has a dependence upon the POD, resulting in the need of a trade-off in order to optimize the model.

Now the importance of the properties of the test samples, as given by τ, in determining the performance of the model become clearer: differently from the POD, the False Alarm Rate is a posterior probability containing the prior knowledge about the class distributions, thus the dependence upon τ arises naturally from the Bayes' theorem. This is also in agreement with our discussion in sec. 3.2 where we found that whereas the FAR contours follow exactly the Equation (23), the POD is essentially independent from τ. In fact, by rewriting the POD as

$$\mathrm{POD} = \frac{\mathrm{TP}}{(n_1^{(total)} - n_1^{(train)})} \tag{30}$$

and the Equation (26) as

$$\frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}} = \frac{\mathrm{FP}}{(\tilde{N} - n_1^{(total)} + n_1^{(train)})}, \tag{31}$$

the Equation (28) yields the contours found in Equation (22).

We remark that in the logistic probability, the prior knowledge about the sample is contained into the constant of the model w_0, being the other coefficients not affected. In order to adjust the model with respect the new prior probability of the occurrence of a SPE associated event, the following correction (King & Zeng, 2001) could be introduced without retraining the model:

$$\tilde{w_0} = w_0 + \log\left[\frac{p(\mathcal{C}_{spe})}{p(\mathcal{C}_{no-spe})}\right]. \tag{32}$$

4 Discussion and Conclusion

In this work we apply the machine learning approach to SPE forecasting and analyse the effects induced by small-sample size and class imbalancing. Following the ESPERTA technique, we used the logistic regression model with three input parameters, the flare heliolongitude, the $1-8$ Å SXR fluence and the ~ 1 MHz radio fluence. We optimize the model weights through basic, weighted and SMOTE MLE. When using the whole dataset of $\geq M2$ flare over the period 1995-2017, we obtained a POD $= 0.76$ and a FAR $= 0.42$ for the basic MLE, whereas no substantial improvement is found by using the weighted and SMOTE MLE. Note that the POD is computed without taking into account the $< M2$ flares associated SPEs, as the goodness of a ML alogrithm, i.e. its ability to learn, has to be evaluated over the ingested dataset. For comparison, we recomputed the ESPERTA scores over the period 1995-2017 in a similar manner from results of Alberti et al., 2017, 2019 and by also considering the corrections of Laurenza et al. 2018 (see their footnotes number 9). The resulting ESPERTA scores are quite comparable with a POD $=$ 78% (73/94) and a FAR $=$ 38% (44/117), although they are derived by including the training sample (covering the 1995-2005 period, Laurenza et al., 2009) in the validation. On the contrary our validation provides scores that are independent on the training and test sample choices.

The present approach is found to be more performant, with POD $= 0.83$ and FAR $= 0.39$, for longitudes of the associated flare $> -20°$, i.e., for central meridian and well-connected SPEs, which are the most hazardous, having the fastest SPE proton onsets, rise times and peak intensities (Kallenrode, 1993; Posner, 2007; Richardson et al., 2014; Papaioannou et al., 2016). On the contrary, SPEs that are not well magnetically connected tend to rise slowly to reduced storm levels and lead to larger advance warning times. Thus, for such events, the radiation risk reduction from any SPE warning system is rather limited, whereas it is most relevant to focus on SPEs that are magnetically well connected (Posner & Strauss, 2020).

We demonstrate that the major drawback in predicting the occurrence of a SPE event in the framework of statistical forecasting, is driven by the optimization of the FAR which, according to Equation (22), depends on the fraction of the events into the sample (i.e., imbalancing) to be predicted. In particular, the greater the imbalancing, the greater the FAR is affected by the presence of a false positive. As a matter of fact, all the SPE forecasting methods present a quite high FAR, generally comprised between 30−55% (Anastasiadis et al., 2017). We explain the high FAR from a Bayesian point of view and show that the FAR explicitly contains the prior knowledge about the class distributions. We point out that this is a critical issue of any statistical approach, and thus the model validation must be done by preserving the class distributions within the training and test datasets.

Recently Núñez and Paul-Pena (2020) obtained POD = 85.3% and FAR = 54.6% using a ML decision tree algorithm validated with the 20-fold method. By comparing their results with POD = 80% and FAR = 45% obtained here through the cross-validated SMOTE MLE over the whole dataset, we observe a quite better FAR despite the fact that we have a greater imbalancing (1 : 9) with respect to Núñez and Paul-Pena (2020) (1.5 : 8.5). This indicates that the logistic model is more suitable than the DT one, given the binary nature of the SPE occurrence forecasting. Moreover, our cross validation is less biased than that in Núñez and Paul-Pena (2020) since it has been performed by: 1) preserving the ratio between SPE and NO SPE events in each fold; 2) using $k = 5$ in order to have a proper statistically significant number of SPE events in each fold.

To sum up, the ML application to SPE forecasting is limited by the small size of the SPE sample with respect to the non SPE associated flares, naturally leading to a greater FAR. We remark the importance of performing the cross validation by preserving the class distributions within the training and test datasets. In order to reduce the high FAR inherent to SPEs forecasting, it should be used features more directly linked to the physical cause of SPE acceleration, supposedly more effective in class separation, instead of associated parameters like those used in this work.

Appendix A SPE catalog 1995 - 2017

In this section we report the complete SPE catalog of events used in the construction of the model. The events are listed in Table A1. All the flares listed here are associated to a SPE observed at the Earth with flux ≥ 10 pfu at > 10 MeV. The features used in the model are the flare longitude, here reported as Hα location, the SXR fluence in J m^{-2} and the Radio fluence in sfu×min computed according to Laurenza et al. (2009). Each SXR fluence value has a correspondent flag in the range $1-7$ that indicates extrapolations as well as ad hoc adjustments involved in the SXR flux integration. We report also the dates of flare SXR peaks and their class. The ~ 1 MHz Radio frequency is listed in the last column .

Acknowledgments

The authors thanks E. W. Cliver for helpful discussions and comments on the manuscript. The authors also thanks the National Oceanic Atmospheric Administration for providing GOES data files and solar flares weakly reports, publicly accessible at `https://ngdc.noaa.gov/stp/space-weather/solar-data/solar-features/solar-flares/h-alpha/tables/` and `https://ftp.swpc.noaa.gov/pub/warehouse/`. Authors thanks also the National Aeronautics and Space Administration for providing Wind/WAVES data available at `https://solar-radio.gsfc.nasa.gov/wind/index.html`. M.S. acknowledge the National Institute for Astrophysics, the University of Rome Tor Vergata" and the University of Rome La Sapienza" for the joint Ph.D. program Astronomy, Astrophysics and Space Science". S.B., M.L., G.C. and M.F.M. acknowledge the financial support by

Table A1. SPE Flare List (1995-2017).

Event Number	SXR Date	SXR Time (hh:mm)	SXR Class	Hα Location	SXR Fluence (J/m^2)	SXR Flag	Radio Fluence (sfu x min)	Radio Frequency (kHz)
1	1995 Oct 20	06:06	M 1.7	S11W53	3.28×10^{-2}	5	5.99×10^{5}	940
2	1997 Nov 04	05:58	X 2.1	S15W34	5.86×10^{-2}	7	1.20×10^{7}	940
3	1997 Nov 06	11:55	X 9.4	S18W63	3.61×10^{-1}	7	1.87×10^{7}	940
4	1998 May 02	13:42	X 1.2	S15W15	7.37×10^{-2}	5	2.14×10^{7}	940
5	1998 May 06	08:09	X 2.8	S11W65	2.35×10^{-1}	5	8.85×10^{6}	940
6	1998 May 09	03:40	M 7.7	W100	1.08×10^{-1}	5	2.69×10^{6}	940
7	1998 Aug 24	22:12	X 1.1	N35E09	1.88×10^{-1}	5	1.79×10^{7}	940
8	1998 Sep 30	13:48	M 3.0	N23W81	9.61×10^{-2}	2	7.09×10^{5}	940
9	1999 Jun 04	07:03	M 4.2	N17W69	2.62×10^{-2}	5	3.95×10^{6}	940
10	2000 Apr 04	15:39	M 1.0	N15W63	3.30×10^{-2}	2	9.24×10^{6}	940
11	2000 Jun 06	15:25	X 2.5	N21E15	4.22×10^{-1}	5	1.28×10^{7}	940
12	2000 Jun 10	17:00	M 5.6	N22W39	1.02×10^{-1}	5	9.57×10^{6}	940
13	2000 Jul 14	10:23	X 6.1	N22W07	1.35	5	1.20×10^{7}	940
14	2000 Jul 22	11:32	M 3.9	N14W56	8.18×10^{-2}	5	1.69×10^{5}	940
15	2000 Sep 12	12:12	M 1.0	S19W08	2.94×10^{-2}	1	5.43×10^{6}	940
16	2000 Oct 16	07:35	M 2.8	W110	8.54×10^{-2}	1	7.18×10^{4}	940
17	2000 Nov 08	23:27	M 7.9	N10W77	3.36×10^{-1}	3	4.51×10^{6}	940
18	2000 Nov 24	15:13	X 2.5	N21W08	1.64×10^{-1}	5	6.77×10^{6}	940
19	2000 Nov 25	01:31	M 8.4	N07E50	2.66×10^{-1}	5	1.69×10^{6}	940
20	2001 Jan 28	15:58	M 1.7	S04W59	3.54×10^{-2}	5	1.60×10^{6}	940
21	2001 Mar 29	10:15	X 1.8	N14W13	2.74×10^{-1}	5	3.83×10^{5}	940
22	2001 Apr 02	21:50	X18.4	N18W82	1.62	5	2.75×10^{6}	940
23	2001 Apr 10	05:26	X 2.3	S23W09	3.66×10^{-1}	5	9.50×10^{6}	940
24	2001 Apr 12	10:28	X 2.2	S19W43	4.02×10^{-1}	5	6.54×10^{6}	940
25	2001 Apr 15	13:50	X15.8	S20W85	6.20×10^{-1}	7	8.77×10^{6}	940
26	2001 May 07	12:20	C 4.1	W35	1.22×10^{-2}	5	1.50×10^{4}	940
27	2001 Aug 09	11:22	C 3.9	W10	1.08×10^{-2}	5	1.57×10^{4}	940
28	2001 Sep 15	11:28	M 1.6	S21W49	5.35×10^{-2}	2	1.98×10^{4}	940
29	2001 Sep 24	10:35	X 2.7	S17E29	1.09	3	1.48×10^{6}	940
30	2001 Oct 01	05:15	M 9.1	S22W85	7.56×10^{-2}	5	1.12×10^{5}	940
31	2001 Oct 19	16:30	X 1.8	N15W29	1.66×10^{-1}	5	3.38×10^{4}	940
32	2001 Oct 22	15:08	M 7.0	S17E19	1.89×10^{-1}	5	1.77×10^{7}	940
33	2001 Nov 04	16:19	X 1.1	N07W19	2.76×10^{-1}	2	1.36×10^{7}	940
34	2001 Nov 17	05:23	M 3.0	S13E42	1.34×10^{-1}	3	3.69×10^{6}	940
35	2001 Nov 22	20:34	M 4.1	S25W67	6.49×10^{-2}	5	4.74×10^{6}	940
36	2001 Nov 22	23:27	X 1.0	S15W34	4.68×10^{-1}	3	1.38×10^{5}	940
37	2001 Dec 26	05:36	M 7.6	N08W54	6.30×10^{-1}	4	1.14×10^{6}	940
38	2001 Dec 28	20:42	X 3.5	S26E95	2.92	4	4.43×10^{6}	940
39	2002 Jan 14	06:23	M 4.8	W90	4.03×10^{-1}	4	9.69×10^{4}	940
40	2002 Feb 20	06:12	M 5.7	N12W72	1.75×10^{-2}	7	7.40×10^{6}	940
41	2002 Mar 15	23:06	M 2.3	S08W03	6.34×10^{-2}	1	2.15×10^{6}	940
42	2002 Mar 18	02:30	M 1.1	W22	1.73×10^{-2}	5	2.67×10^{5}	940
43	2002 Apr 17	08:24	M 2.9	S14W36	1.35×10^{-1}	3	6.93×10^{5}	940
44	2002 Apr 21	01:47	X 1.7	S14W84	7.82×10^{-1}	3	4.51×10^{6}	940
45	2002 May 22	03:48	C 5.2	S22W53	1.82×10^{-2}	1	2.02×10^{6}	940
46	2002 Jul 15	20:08	X 3.2	N19W01	1.49×10^{-1}	7	9.81×10^{6}	940

Table 1. (continued)

Event Number	SXR Date	SXR Time (hh:mm)	SXR Class	Hα Location	SXR Fluence (J/m^2)	SXR Flag	Radio Fluence (sfu x min)	Radio Frequency (kHz)
47	2002 Jul 20	21:28	X 3.4	E100	1.08	5	3.32×10^{6}	940
48	2002 Aug 14	02:11	M 2.6	N10W54	1.06×10^{-1}	3	9.51×10^{5}	940
49	2002 Aug 22	01:57	M 5.9	S07W62	3.82×10^{-2}	5	1.02×10^{7}	940
50	2002 Aug 24	01:11	X 3.5	S02W81	5.75×10^{-1}	5	7.23×10^{5}	940
51	2002 Sep 05	17:04	C 5.2	N09E28	2.49×10^{-2}	3	2.34×10^{5}	940
52	2002 Nov 09	13:23	M 4.9	S12W29	5.52×10^{-2}	5	8.14×10^{6}	940
53	2003 May 28	00:27	X 3.9	S07W21	3.12×10^{-1}	5	7.20×10^{6}	940
54	2003 May 31	02:24	X 1.0	S07W65	1.20×10^{-1}	5	7.96×10^{6}	940
55	2003 Oct 26	18:11	X 1.4	N02W38	3.83×10^{-1}	1	1.43×10^{6}	940
56	2003 Oct 28	11:10	X18.4	S16E07	1.96	5	2.16×10^{7}	940
57	2003 Oct 29	20:49	X10.8	S15W02	9.80×10^{-1}	5	8.79×10^{6}	940
58	2003 Nov 02	17:25	X 9.3	S14W56	1.09	5	2.70×10^{6}	940
59	2003 Nov 04	19:44	X18.4	S19W83	2.65	1	9.53×10^{5}	940
60	2003 Nov 20	23:53	M 6.2	N02W17	2.82×10^{-2}	7	7.07×10^{6}	940
61	2004 Apr 11	04:19	M 1.0	S14W47	1.72×10^{-2}	5	3.03×10^{6}	940
62	2004 Jul 25	15:15	M 1.2	N08W33	3.25×10^{-2}	1	7.51×10^{4}	940
63	2004 Sep 19	17:11	M 2.0	N05W58	5.46×10^{-2}	5	4.20×10^{6}	940
64	2004 Nov 07	16:06	X 2.2	N09W17	2.08×10^{-1}	5	1.36×10^{6}	940
65	2004 Nov 10	02:13	X 2.8	N09W49	1.68×10^{-1}	7	1.84×10^{6}	940
66	2005 Jan 15	23:00	X 2.9	N14W08	8.63×10^{-1}	2	1.01×10^{6}	940
67	2005 Jan 17	09:52	X 4.2	N14W24	7.20×10^{-1}	5	1.63×10^{6}	940
68	2005 Jan 20	07:00	X 7.9	N12W58	1.97	5	1.66×10^{7}	940
69	2005 May 13	16:57	M 8.5	N12E11	2.50×10^{-1}	5	1.79×10^{7}	940
70	2005 Jun 16	20:22	M 4.3	N09W87	7.75×10^{-2}	5	6.94×10^{5}	940
71	2005 Jul 13	14:49	M 5.6	N13W75	4.64×10^{-1}	4	1.08×10^{5}	940
72	2005 Jul 14	10:54	X 1.4	W95	6.63×10^{-1}	3	2.65×10^{4}	940
73	2005 Jul 27	05:01	M 3.8	<E90	1.16×10^{-1}	5	8.83×10^{4}	940
74	2005 Aug 22	17:28	M 6.2	S12W60	2.87×10^{-1}	3	1.54×10^{6}	940
75	2005 Sep 07	17:40	X18.1	S06E89	6.65	3	1.42×10^{7}	940
76	2005 Sep 13	20:04	X 1.6	S09E05	4.86×10^{-1}	5	1.49×10^{5}	940
77	2006 Dec 05	10:35	X 9.0	S07E79	6.12×10^{-1}	5	1.90×10^{6}	916
78	2006 Dec 13	02:39	X 3.4	S05W23	5.88×10^{-1}	5	1.82×10^{7}	916
79	2006 Dec 14	22:15	X 1.5	S06W46	1.36×10^{-1}		7.52×10^{6}	
80	2010 Aug 14	10:05	C 4.4	N17W52	1.19×10^{-2}		1.29×10^{5}	916
81	2011 Mar 08	10:44	M 5.3	S17W86	3.98×10^{-2}	2	5.55×10^{3}	916
82	2011 Jun 07	06:41	M 2.5	S21W64	4.91×10^{-2}	5	1.80×10^{7}	916
83	2011 Aug 04	03:57	M 9.3	N15W49	6.07×10^{-2}	5	8.78×10^{6}	916
84	2011 Aug 09	08:05	X 6.9	N17W83	1.77×10^{-1}	7	5.71×10^{6}	916
85	2011 Sep 22	11:01	X 1.4	N11E74	4.78×10^{-1}	2	4.32×10^{6}	916
86	2011 Nov 26	07:10	C 1.2	N08W49	1.47×10^{-2}		1.74×10^{5}	916
87	2012 Jan 23	03:59	M 8.7	N28W36	3.97×10^{-2}	5	5.26×10^{5}	916
88	2012 Jan 27	18:37	X 1.7	N27W71	2.33×10^{-1}	5	4.38×10^{6}	916
89	2012 Mar 07	00:24	X 5.4	N17E15	6.89×10^{-1}	5	2.19×10^{7}	916
90	2012 Mar 13	17:41	M 7.9	N18W62	2.65×10^{-1}	3	2.92×10^{6}	916
91	2012 May 17	01:47	M 5.1	N12W89	1.21×10^{-1}	5	9.08×10^{6}	916
92	2012 Jun 14	14:35	M 1.9	S17E14	1.55×10^{-1}		9.52×10^{4}	

Table 1. (continued)

Event Number	SXR Date	SXR Time (hh:mm)	SXR Class	Hα Location	SXR Fluence (J/m^2)	SXR Flag	Radio Fluence (sfu x min)	Radio Frequency (kHz)
93	2012 Jul 06	01:40	M 2.9	S18W41	4.62×10^{-2}	5	1.21×10^{7}	916
94	2012 Jul 12	16:49	X 1.4	S16W09	5.28×10^{-1}	3	7.54×10^{5}	916
95	2012 Jul 17	17:15	M 1.7	S17W75	1.86×10^{-1}		3.27×10^{5}	916
96	2012 Jul 19	05:58	M 7.7	W99	3.58×10^{-1}		4.88×10^{5}	
97	2012 Aug 31	20:43	C 8.4	S06E20	6.57×10^{-2}		3.19×10^{6}	916
98	2012 Sep 27	23:57	C 3.7	N08W41	4.19×10^{-3}		6.18×10^{4}	916
99	2013 Mar 15	06:58	M 1.1	N11E12	6.64×10^{-2}		3.92×10^{4}	
100	2013 Apr 11	07:16	M 6.5	N09E12	7.11×10^{-2}	5	3.38×10^{7}	916
101	2013 May 15	01:48	X 1.2	N11E51	1.19×10^{-1}	5	1.58×10^{4}	916
102	2013 May 22	13:32	M 5.0	N15W70	1.77×10^{-1}	3	5.74×10^{5}	916
103	2013 Jun 21	03:14	M 2.9	S16E66	8.11×10^{-2}	2	6.18×10^{4}	916
104	2013 Sep 29	23:37	C 1.2	N15W40	3.07×10^{-3}		6.94×10^{4}	916
105	2013 Dec 28	18:02	C 9.3	S18E07	4.80×10^{-3}		1.23×10^{4}	916
106	2014 Jan 07	10:13	M 7.2	S13E11	2.95×10^{-2}	5	1.75×10^{7}	916
107	2014 Feb 20	07:55	M 3.0	S15W67	7.38×10^{-2}	3	1.75×10^{6}	916
108	2014 Feb 25	00:49	X 4.9	S12E82	4.64×10^{-1}	5	6.83×10^{6}	916
109	2014 Apr 18	13:03	M 7.3	S16W41	1.13×10^{-1}	5	7.98×10^{6}	916
110	2014 Sep 10	17:45	X 1.6	N16W06	3.88×10^{-1}	5	3.49×10^{7}	916
111	2015 Jun 21	01:42	M 2.0	N12E16	1.61×10^{-1}		7.82×10^{6}	
112	2015 Jun 25	08:16	M 7.9	N09W42	2.11×10^{-1}		3.51×10^{5}	
113	2016 Jan 02	00:11	M 2.3	S12W27	8.51×10^{-2}		4.53×10^{3}	

Italian MIUR-PRIN grant 2017APKP7T on Circumterrestrial Environment: Impact of Sun-Earth Interaction.

References

Alberti, T., Laurenza, M., Cliver, E., Storini, M., Consolini, G., & Lepreti, F. (2017). Solar activity from 2006 to 2014 and short-term forecasts of solar proton events using the esperta model. *The Astrophysical Journal*, *838*(1), 59.

Anastasiadis, A., Papaioannou, A., Sandberg, I., Georgoulis, M., Tziotziou, K., Kouloumvakos, A., & Jiggens, P. (2017). Predicting flares and solar energetic particle events: The forspef tool. *Solar Physics*, *292*(9), 1–21.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer. Retrieved from `http://research.microsoft.com/en-us/um/people/cmbishop/prml/`

Boser, B. E., Guyon, I. M., & Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on computational learning theory* (pp. 144–152).

Boubrahimi, S. F., Aydin, B., Martens, P., & Angryk, R. (2017). On the prediction of¿ 100 mev solar energetic particle events using goes satellite data. In *2017 ieee international conference on big data (big data)* (pp. 2533–2542).

Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5–32.

Browne, M. W. (2000). Cross-validation methods. *Journal of mathematical psychology*, *44*(1), 108–132.

Butler, K. T., Davies, D. W., Cartwright, H., Isayev, O., & Walsh, A. (2018). Machine learning for molecular and materials science. *Nature*, *559*(7715), 547–555.

Camporeale, E., Wing, S., & Johnson, J. (2018). *Machine learning techniques for space weather*. Elsevier.

Carleo, G., Cirac, I., Cranmer, K., Daudet, L., Schuld, M., Tishby, N., ... Zdeborová, L. (2019). Machine learning and the physical sciences. *Reviews of Modern Physics*, *91*(4), 045002.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, *16*, 321–357.

Collett, D. (2002). *Modelling binary data*. CRC press.

Cox, D. R. (1958). The regression analysis of binary sequences. *Journal of the Royal Statistical Society: Series B (Methodological)*, *20*(2), 215–232.

Cramer, J. S. (2002). *The origins of logistic regression*. Tinbergen Institute Working Paper.

Dietterich, T. (1995). Overfitting and undercomputing in machine learning. *ACM computing surveys (CSUR)*, *27*(3), 326–327.

Fix, E. (1985). *Discriminatory analysis: nonparametric discrimination, consistency properties* (Vol. 1). USAF school of Aviation Medicine.

Friedman, J., Hastie, T., Tibshirani, R., et al. (2001). *The elements of statistical learning* (Vol. 1) (No. 10). Springer series in statistics New York.

Gao, S., & Shen, J. (2007). Asymptotic properties of a double penalized maximum likelihood estimator in logistic regression. *Statistics & probability letters*, *77*(9), 925–930.

Grinstead, C. M., & Snell, J. L. (2012). *Introduction to probability*. American Mathematical Soc.

Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression* (Vol. 398). John Wiley & Sons.

Kahler, S., Cliver, E., & Ling, A. (2007). Validating the proton prediction system (pps). *Journal of atmospheric and solar-terrestrial physics*, *69*(1-2), 43–49.

Kallenrode, M.-B. (1993). Neutral lines and azimuthal transport of solar energetic particles. *Journal of Geophysical Research: Space Physics*, *98*(A4), 5573–5591.

King, G., & Zeng, L. (2001, Spring). Logistic regression in rare events data. *Political Analysis*, *9*, 137–163.

Laurenza, M., Alberti, T., & Cliver, E. (2018). A short-term esperta-based forecast tool for moderate-to-extreme solar proton events. *The Astrophysical Journal*, *857*(2), 107.

Laurenza, M., Cliver, E., Hewitt, J., Storini, M., Ling, A., Balch, C., & Kaiser, M. (2009). A technique for short-term warning of solar energetic particle events based on flare location, flare size, and evidence of particle escape. *Space Weather*, *7*(4).

Laurenza, M., Hewitt, J., Storini, M., et al. (2007). *Solar energetic proton events and soft x-ray flares, 20th ecrs proceedings.*

Lichtenstein, S., Fischhoff, B., & Phillips, L. D. (1977). Calibration of probabilities: The state of the art. *Decision making and change in human affairs*, 275–324.

Maalouf, M. (2011, 07). Logistic regression in data analysis: An overview. *International Journal of Data Analysis Techniques and Strategies*, *3*, 281-299. doi: 10.1504/IJDATS.2011.041335

Malandraki, O. E., & Crosby, N. B. (2018). *Solar particle radiation storms forecasting and analysis: The hesperia horizon 2020 project and beyond*. Springer Nature.

Manski, C. F., & Lerman, S. R. (1977). The estimation of choice probabilities from choice based samples. *Econometrica: Journal of the Econometric Society*, 1977–1988.

McCullagh, P., & Nelder, J. (1989). *Generalized linear models, second edition*. Chapman & Hall. Retrieved from `http://books.google.com/books?id=h9kFH2\_FfBkC`

Niculescu-Mizil, A., & Caruana, R. (2005). Predicting good probabilities with supervised learning. In *Proceedings of the 22nd international conference on machine learning* (pp. 625–632).

Núñez, M. (2018). Predicting well-connected sep events from observations of solar soft x-rays and near-relativistic electrons. *Journal of Space Weather and Space Climate*, *8*, A36.

Núñez, M., & Paul-Pena, D. (2020). Predicting¿ 10 mev sep events from solar flare and radio burst data. *Universe*, *6*(10), 161.

Papaioannou, A., Sandberg, I., Anastasiadis, A., Kouloumvakos, A., Georgoulis, M. K., Tziotziou, K., ... Hilgers, A. (2016). Solar flares, coronal mass ejections and solar energetic particle event characteristics. *Journal of Space Weather and Space Climate*, *6*, A42.

Posner, A. (2007). Up to 1-hour forecasting of radiation hazards from solar energetic ion events with relativistic electrons. *Space Weather*, *5*(5).

Posner, A., & Strauss, R. (2020). Warning time analysis from sep simulations of a two-tier release system applied to mars exploration. *Space Weather*, *18*(4), e2019SW002354.

Quinlan, J. R. (1996). Learning decision tree classifiers. *ACM Computing Surveys (CSUR)*, *28*(1), 71–72.

Richardson, I., von Rosenvinge, T., Cane, H., Christian, E., Cohen, C., Labrador, A., ... Stone, E. (2014). ¿ 25 mev proton events observed by the high energy telescopes on the stereo a and b spacecraft and/or at earth during the first seven years of the stereo mission. In *Coronal magnetometry* (pp. 437–485). Springer.

Smart, D., & Shea, M. (1979). Pps 76-a computerized" event mode. In *Solar-terrestrial predictions proceedings* (Vol. 1, p. 406).

Smart, D., & Shea, M. (1989). Pps-87: a new event oriented solar proton prediction model. *Advances in Space Research*, *9*(10), 281–284.

St. Cyr, O., Posner, A., & Burkepile, J. (2017). Solar energetic particle warnings from a coronagraph. *Space Weather*, *15*(1), 240–257.

Stone, M. (1978). Cross-validation: A review. *Statistics: A Journal of Theoretical and Applied Statistics*, *9*(1), 127–139.

Winter, L. M., & Ledbetter, K. (2015). Type ii and type iii radio bursts and their correlation with solar energetic proton events. *The Astrophysical Journal*, *809*(1), 105.

Figure 1.

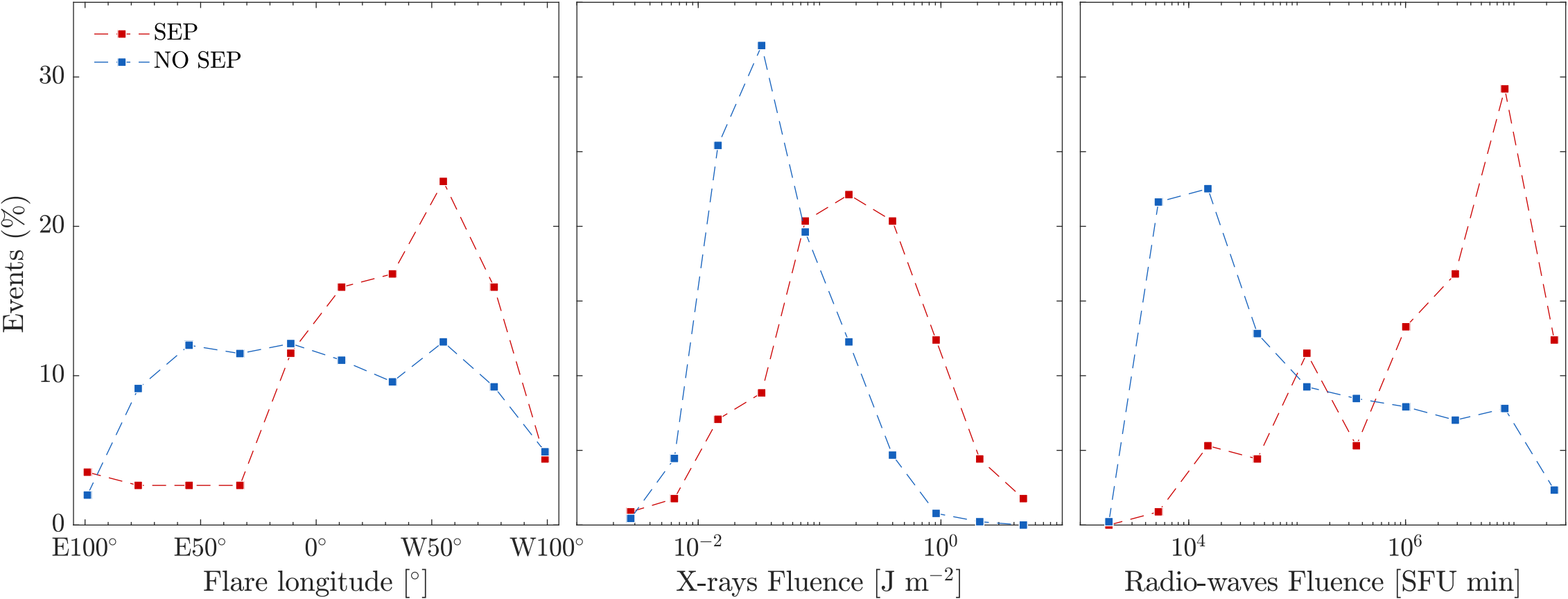

SEP
NO SEP
Events (%)
0
10
20
30
E100°
E50°
0°
W50°
W100°
Flare longitude [°]
10^{-2}
10^{0}
X-rays Fluence [J m^{-2}]
10^{4}
10^{6}
Radio-waves Fluence [SFU min]

Figure 2.

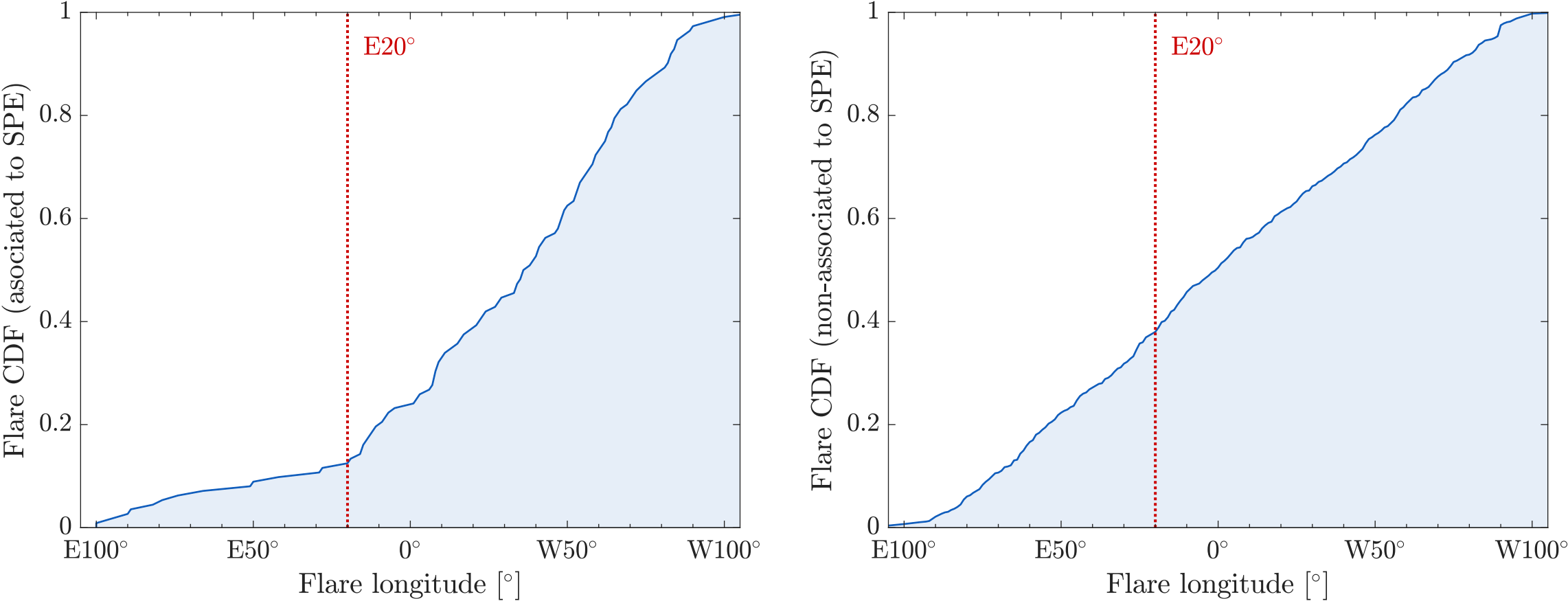
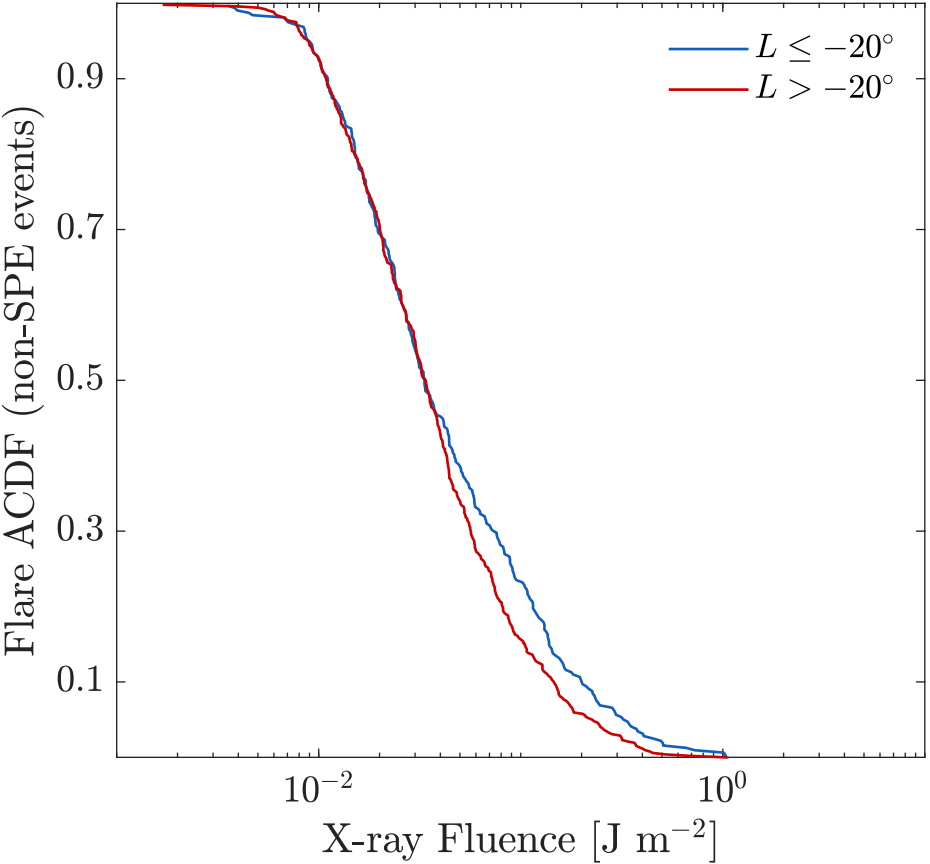

E20°
Flare CDF (asociated to SPE)
Flare longitude [°]
E100°
E50°
0°
W50°
W100°
0
0.2
0.4
0.6
0.8
1
E20°
Flare CDF (non-associated to SPE)
Flare longitude [°]
E100°
E50°
0°
W50°
W100°
0
0.2
0.4
0.6
0.8
1

Figure 3.


$L \leq -20°$
$L > -20°$
0.9
0.7
0.5
0.3
0.1
Flare ACDF (non-SPE events)
10^{-2}
10^{0}
X-ray Fluence [J m^{-2}]

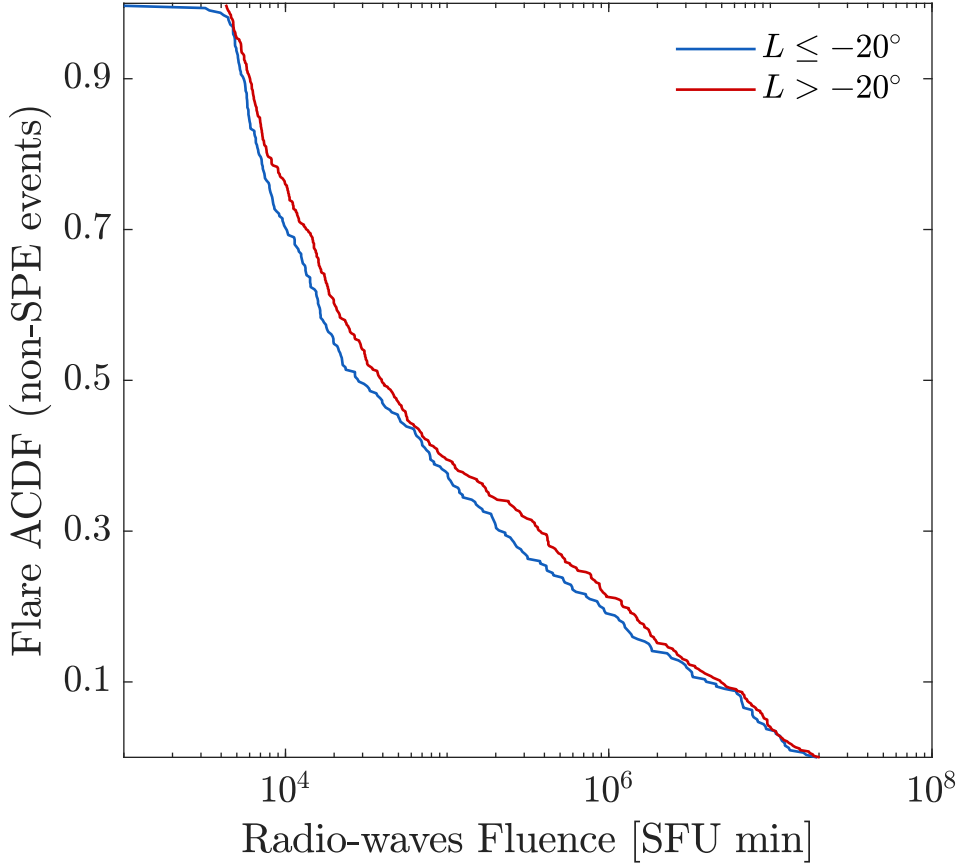

$L \leq -20°$
$L > -20°$
0.9
0.7
0.5
0.3
0.1
Flare ACDF (non-SPE events)
10^{4}
10^{6}
10^{8}
Radio-waves Fluence [SFU min]

Figure 4.

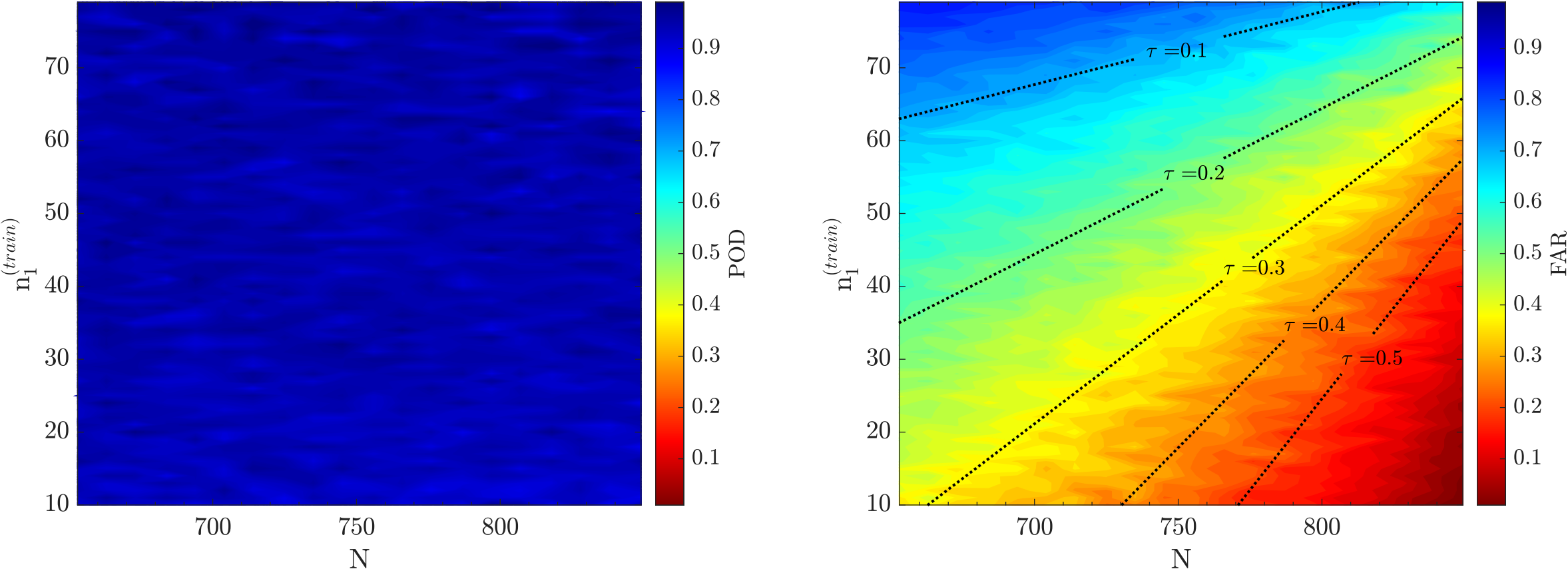
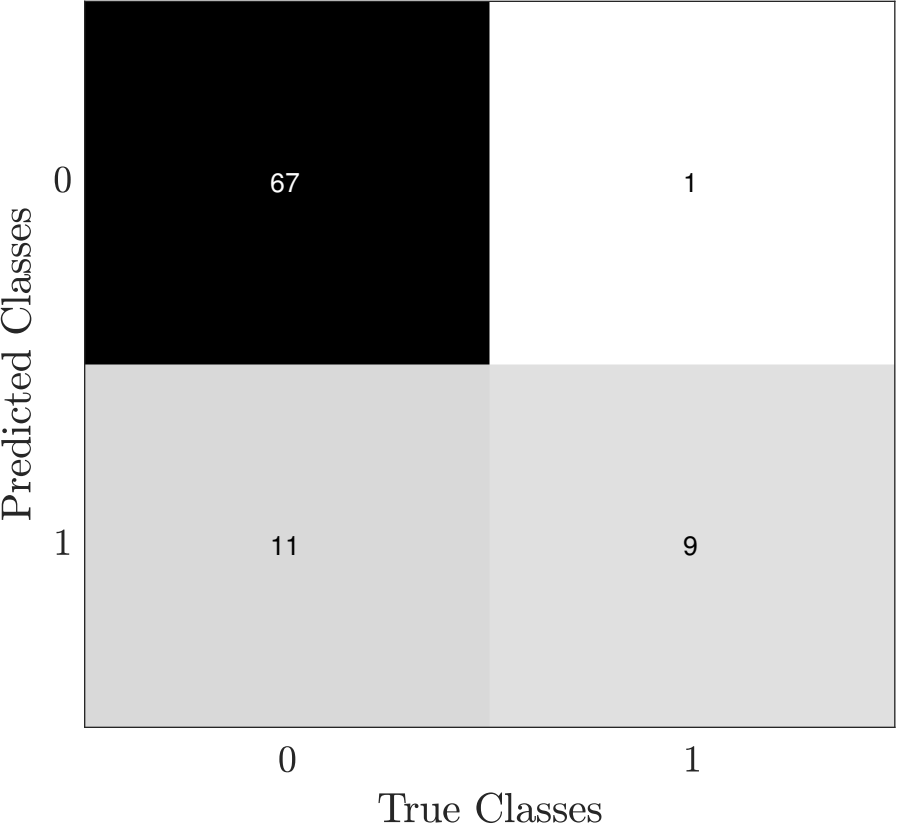

70
60
50
40
30
20
10
$n_1^{(train)}$
700
750
800
N
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
POD
τ =0.1
τ =0.2
τ =0.3
τ =0.4
τ =0.5
FAR

Figure 5.


Predicted Classes
0
1
67
1
11
9
0
1
True Classes

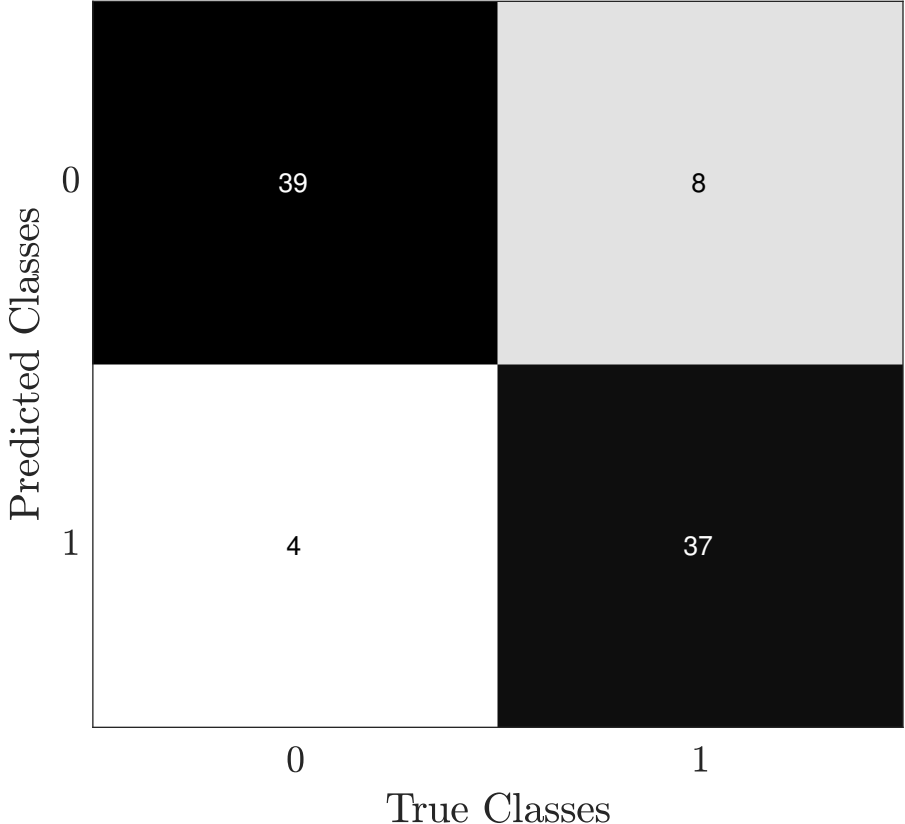

Predicted Classes
0
1
39
8
4
37
0
1
True Classes

Figure 6.

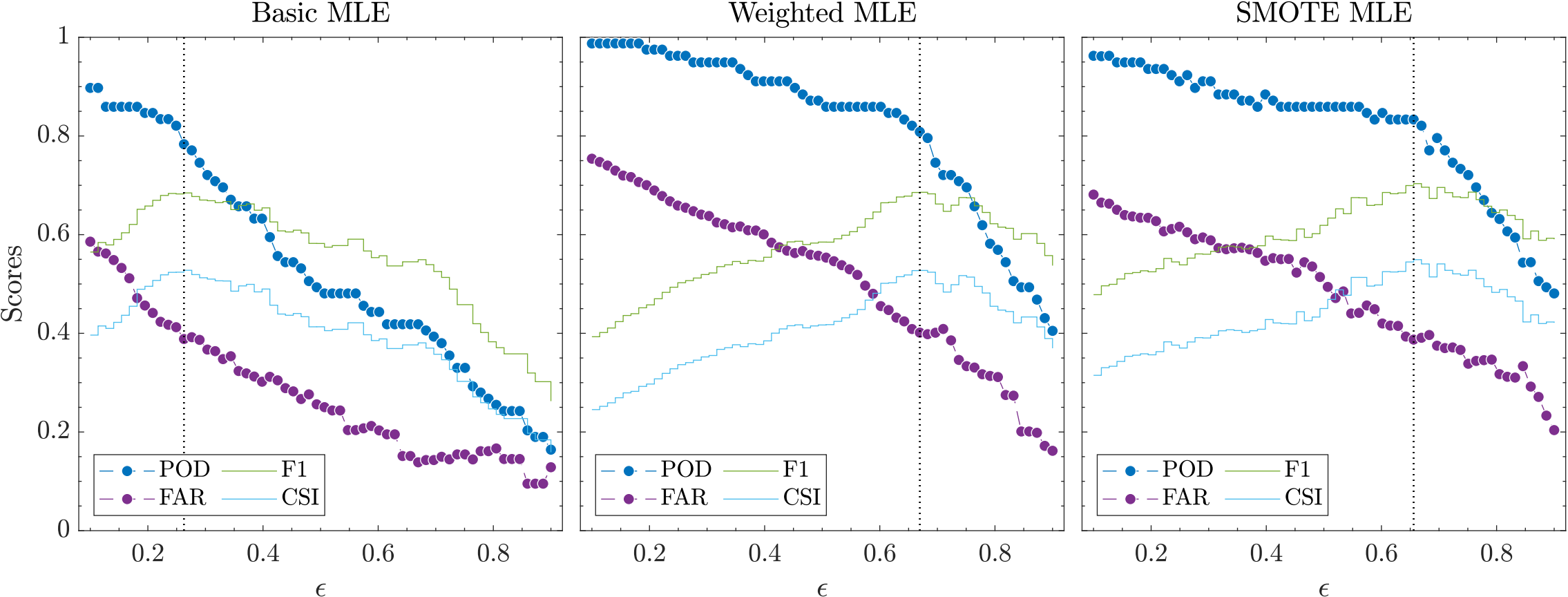
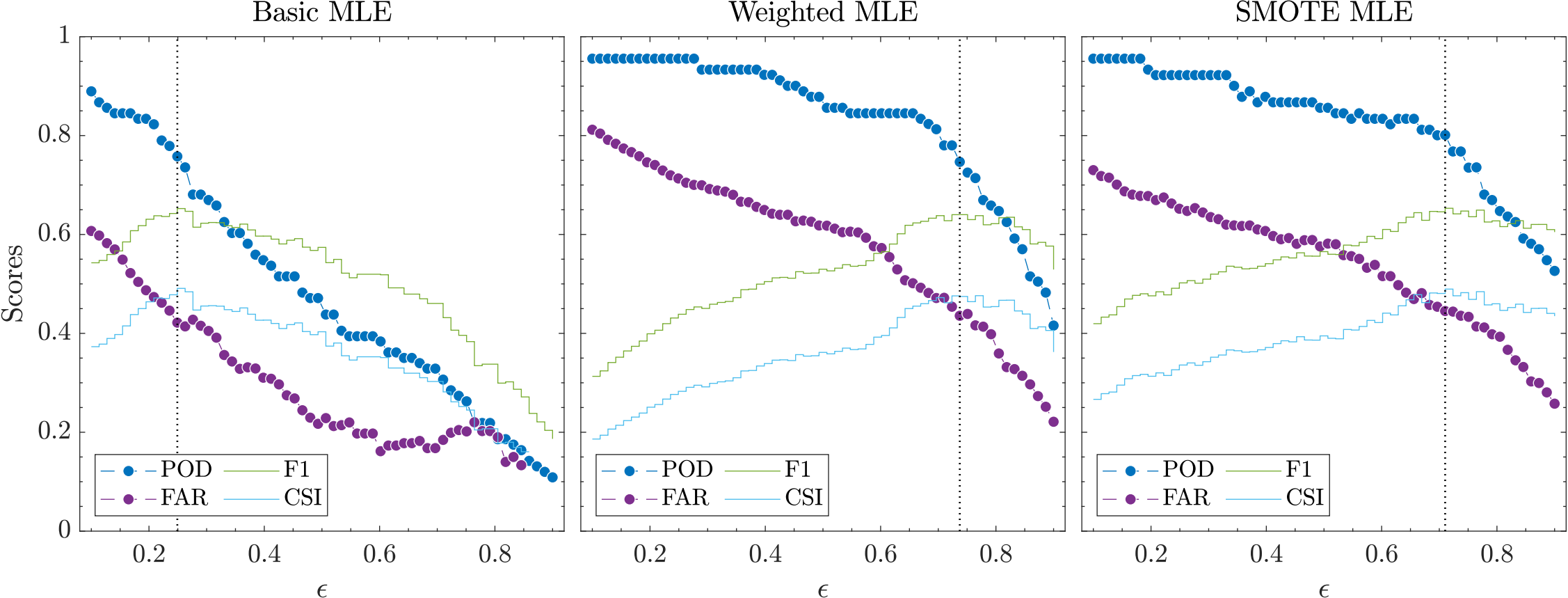

Basic MLE
Weighted MLE
SMOTE MLE
Scores
1
0.8
0.6
0.4
0.2
0
0.2
0.4
0.6
0.8
ϵ
POD
FAR
F1
CSI

Figure 7.


Basic MLE
Weighted MLE
SMOTE MLE
Scores
1
0.8
0.6
0.4
0.2
0
0.2
0.4
0.6
0.8
ϵ
POD
FAR
F1
CSI