Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Probabilistic forecasting of >10 MeV solar proton events from flare, CME and radio observations of solar cycles 23–25: physical predictors and verification of tree-ensemble classifiers

Igor Davydenkov

davydenkov.ii@phystech.edu

Space Research Institute

Vadim Ozheredov

Space Research Institute

Research Article

Keywords:

Posted Date: November 17th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11606790/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Probabilistic forecasting of >10 MeV solar proton events from flare, CME and radio observations of solar cycles 23–25: physical predictors and verification of tree-ensemble classifiers

I. I. Davydenkov*, V. A. Ozheredov

Space Research Institute of the Russian Academy of Sciences, Profsoyuznaya 84/32, Moscow, 117997, Russia

I. I. Davydenkov: ORCID 0009-0008-7369-5831, davydenkov.ii@phystech.edu

V. A. Ozheredov: ORCID 0000-0001-6641-444X, ozheredov2016@gmail.com

Abstract

We study the probability that a strong solar flare accompanied by a coronal mass ejection (CME) produces a >10 MeV proton event at the Earth. The sample contains 964 flares of GOES class M2.0 and above observed in 1997–2022 with associated CMEs and radio data; 94 of them are parent flares of NOAA-listed proton events (≥10 pfu). The sample was cleaned of duplicate flare records and erroneous flare coordinates, and 36 predictors were computed from soft X-ray, CME, type II/IV radio, background-subtracted microwave and active-region data. Gradient boosting (CatBoost), Random Forest and a six-parameter logistic model were verified by repeated nested cross-validation and by an independent test on 2011–2022. CatBoost reaches an area under the receiver operating characteristic curve (ROC AUC) of 0.935 ± 0.004 and a Brier skill score of 0.33; at the threshold maximising the Heidke skill score it warns 72% of the proton-producing eruptions with a false-alarm ratio of 0.51. The skill is preserved for solar cycles 24–25 (ROC AUC 0.94, true skill statistic 0.72). The information is carried mainly by type II radio bursts, source longitude, soft X-ray fluence and CME speed; longitude acts as a modulating condition with an effective boundary near E20, and the contribution of the CME speed changes sign near 1000 km s^{-1}. The logistic model reaches ROC AUC 0.92. About a third of the persistent false alarms in 2010–2022 were followed by sub-threshold proton enhancements. Heuristic event labels and duplicate records reduce the apparent ROC AUC to 0.83–0.85.

Keywords: Sun: flares – Sun: coronal mass ejections – Sun: particle emission – Sun: radio radiation – space weather – methods: statistical

1 Introduction

Large solar energetic particle (SEP) events are produced by two acceleration processes that usually operate together: magnetic reconnection in the flare and diffusive acceleration at the shock driven by a fast coronal mass ejection (CME) [1, 2, 3]. Protons above 10 MeV reach the Earth promptly only if the acceleration region is magnetically connected to the observer through the Parker spiral [4]; for a solar-wind speed of ≈ 400 km s^{-1} the nominal footpoint lies at western heliolongitudes of ≈ 50°–60°. The peak intensity of an event therefore depends on the energy released in the flare, on the speed and width of the CME, on the presence of a shock in the low corona (radio type II emission), on the population of seed particles, and on the source longitude [5, 6, 7, 8, 9]. Because each of these quantities is measured only indirectly, the relation

*Corresponding author: I. I. Davydenkov, davydenkov.ii@phystech.edu

between the solar observables and the SEP outcome is statistical, and the large majority of M- and X-class flares, even those accompanied by fast CMEs, are not followed by a >10 MeV event at the Earth.

Operational and research forecasting schemes exploit this physics in different ways. Empirical rules based on soft X-ray (SXR) and radio signatures of particle escape [10, 11, 12], regression on flare and CME parameters [13, 14], and supervised machine learning [15, 16] have been compared in the review [17]. Recently AbuBakr Ali et al. [16] combined flare, CME and radio-burst parameters in a Random Forest classifier and reported a probability of detection (POD) of 0.85 with a false-alarm ratio (FAR) of 0.30 for events with sweep-frequency radio data.

The verification of such classifiers is delicate. The event sample is small (typically $\sim 10^2$ positive cases per solar cycle), the classes are strongly imbalanced, and categorical scores such as POD and FAR depend on the alert threshold. If the threshold is tuned on the same data that are used to report the scores, or if the same flare enters both the training and the test set, the skill is overestimated [18]. The skill also depends on how the SEP label is attached to a flare: the parent flare of a proton event is not always the largest flare in the preceding day.

In this paper we study the probability that an M2.0-class or stronger flare with an associated CME produces a >10 MeV proton event at the Earth, using 26 years of observations (1997–2022, solar cycles 23–25). Compared with the previous version of this study, we (i) rebuilt the event sample on the NOAA GOES flare list, removing duplicated flare records and erroneous flare coordinates, (ii) attached the SEP labels using the parent-flare associations of the NOAA Space Weather Prediction Center (SWPC) proton-event list, (iii) recomputed the microwave and active-region predictors, and (iv) verified the classifiers with repeated nested cross-validation and with an independent chronological test. The aim is to establish which physical predictors carry the forecast information, what skill a flexible tree ensemble adds to a simple physical baseline, and how sensitive the reported skill is to the treatment of the data. All data and code are publicly available (Sect. 7).

2 Observational data

2.1 Event sample

The initial sample was compiled from the Heliophysics Event Knowledgebase (HEK) [19] and contains HEK flare records of GOES class ≥M2.0 with peak times between 1997 January 1 and 2022 December 31 that satisfy two observational conditions: (a) a CME from the SOHO/LASCO catalogue [20] first appears between 2 h before and 8 h after the SXR peak (when several CMEs occur in this interval the fastest one is taken), and (b) either sweep-frequency radio evidence exists (a ground-based type II or type IV burst, or a WIND/WAVES [21] type II burst from the CDAW list [22]), or the Radio Solar Telescope Network (RSTN) recorded the flare interval at one or more of the fixed frequencies 245, 410, 610, 1415, 2695, 4995, 8800 and 15 400 MHz. The selection follows the multi-source design of [16] but uses no synthetic (oversampled) events. The initial table contains 1150 HEK records.

2.2 Identification of SEP events

A flare is labelled SEP-productive ($y = 1$) if it is the parent flare of a proton event in the NOAA SWPC list of "Solar Proton Events Affecting the Earth Environment" [33], i.e. an event in which the GOES >10 MeV integral flux exceeds 10 pfu (1 pfu = 1 particle $\mathrm{cm^{-2}\,s^{-1}\,sr^{-1}}$) for at least three consecutive 5-min intervals. The list gives, for each proton event, the time of maximum and the importance of the associated flare as identified by SWPC forecasters. Two versions of the list are in use: the preliminary listing mirrored by the NASA Solar Data Analysis Center (SDAC; 1976–2017) and the version distributed by NOAA NCEI (1976–present, retrieved 2026 May 9). We found that each version omits events present in the other (11 and 10 events before 2017 October, respectively; e.g. the ground-level enhancement of 2001 April 15 is missing from the NCEI version, and the events of 2003 October 29 and 2004 November 10 from the SDAC listing), and used their union. Entries with onsets within 3 h, or within 24 h and with the same peak flux or parent flare, were treated as one event. The parent flares were matched to the NOAA GOES flare list (Sect. 2.3) requiring the same GOES letter class and a difference of the times of maximum not exceeding 30 min. Between 1997 and 2022 the merged list contains 150 proton events; 136 of them have an associated flare, 134 of these flares were found in the GOES list, and for 101 events the parent flare is of class ≥M2.0. The parent flares of 94 of these 101 events are in our sample; the remaining seven occurred during the SOHO interruptions of 1998 June–October and 1998 December–1999 February, or are far-eastern and limb events (2005 June 16, 2005 September 7, 2006 December 5) and the X1.3 flare of 2022 March 30, which lack a catalogued CME or radio association in the initial table. The positive class contains 94 flares (some flares are parents of two events) with peak >10 MeV fluxes from 11 to 31 700 pfu (median 111 pfu). The SEP onset follows the SXR peak by a median of 3.1 h (interquartile range 0.8–11.3 h).

2.3 Quality control

Three problems of the initial table were found and corrected. Their effect on the classification skill is quantified in Sect. 5.6.

Duplicate flare records. HEK stores the same flare as separate records produced by different feature-recognition modules (NOAA SEC/SWPC, Yohkoh, SolarSoft Latest Events). Such records share all CME and radio predictors but, in the initial table, could carry different SEP labels. Each record was therefore anchored to the NOAA GOES flare list – the NCEI GOES XRS flare reports [34] up to mid-2017 and the SWPC "Edited Events" reports afterwards – by requiring $|\Delta t_{\mathrm{peak}}| \leq 20$ min and a GOES class within a factor of two; records anchored to the same NOAA flare (or, if unmatched, peaking within 15 min with a similar class) were merged. The merge removed 185 duplicates; one further flare has class below M2.0 in the NOAA list and was discarded. The final sample contains 964 flares (94 SEP, 870 non-SEP; event rate 9.8%).

Flare coordinates. In 229 initial records the "longitude" exceeded 90° in absolute value: for Yohkoh records HEK provides helioprojective coordinates in arcseconds, which had been used as heliographic degrees. Further 306 records had a zero longitude, which in HEK denotes a missing position (e.g. the X8.2 flare of 2017 September 10 at S08W88). We took the heliographic position from the NOAA flare report (658 flares); otherwise from the NOAA Solar Region Summary

(SRS) position of the source active region, rotated to the flare time at $13.2^\circ\,\mathrm{day}^{-1}$ (136 flares); otherwise from SolarSoft records given in degrees (58 flares). For 112 flares (12%), mostly limb or unattributed events, the longitude remains unknown and is treated as missing.

SEP labels. The labels of the previous version were assigned by a heuristic that marks the strongest flare in the 24 h preceding a proton onset. This rule disagrees with the SWPC associations for 38 of the 964 flares (25 SWPC parent flares were not labelled, 13 labelled flares are not SWPC parents). It missed, for example, the X14.4 flare of 2001 April 15 (ground-level enhancement, 951 pfu), the X20 flare of 2001 April 2 and the M9.9 flare of 2001 November 22 (18 900 pfu), and it marked flares that occurred after the parent eruption of an ongoing event.

3 Physical predictors

All predictors are computed from measured or catalogued quantities; missing values are kept missing. The 36 predictors are organised in six physical groups (Table 1).

Flare soft X-rays. With $F_L(t)$ and $F_S(t)$ the GOES 1–8 Å and 0.5–4 Å fluxes and t_0, t_p, t_1 the flare start, peak and end, we use the peak flux $F_\mathrm{peak} = F_L(t_\mathrm{p})$, the rise time $t_\mathrm{p} - t_0$, the duration $t_1 - t_0$ and the SXR fluence

$$\Phi_\mathrm{SXR} = \int_{t_0}^{t_1} F_L(t)\,dt, \tag{1}$$

taken from the integrated 1–8 Å flux reported by NOAA (931 flares; for the other 33 we use $F_\mathrm{peak}(t_1 - t_0)$). The fluence measures the total thermal energy release and is a better proxy of eruption size than the peak flux [5, 8]. The spectral hardness and its evolution are characterised by the ratio $r(t) = F_S(t)/F_L(t)$ during the first ten minutes of the flare,

$$R_\mathrm{XS,10} = r(t_0 + 10\,\mathrm{min}), \qquad D_\mathrm{XS} = \frac{d\hat{r}}{dt}\Big|_{[t_0,\,t_0+10\,\mathrm{min}]}, \qquad D_\mathrm{SXR} = \max_{t_0 \le t \le t_0+10\,\mathrm{min}} \frac{\Delta r}{\Delta t}, \tag{2}$$

where $\hat{r}$ is a linear least-squares fit and $\Delta r/\Delta t$ is the one-minute increment. The ratio r is a proxy of the plasma temperature, and its early growth rate a proxy of impulsive (Neupert-like) heating [23]; these three quantities are available for the 609 flares with GOES 1-min data in the cache.

Magnetic connection. The heliographic longitude λ (positive to the west) and the connection proxy

$$C_\mathrm{W45} = \cos(\lambda - 45^\circ) \tag{3}$$

describe the proximity of the source to the nominal footpoint of the Earth-connected Parker spiral [4].

CME kinematics. We use the plane-of-sky linear speed v_CME and angular width w_CME from the LASCO catalogue [20] and the combination

$$S_\mathrm{CME} = \ln(1 + v_\mathrm{CME})\,\ln(1 + w_\mathrm{CME}), \tag{4}$$

(v in $\mathrm{km\,s^{-1}}$, w in degrees), which increases monotonically with both the shock strength and its angular extent [7, 24].

Sweep-frequency radio. Ground-based type II and type IV flags, the type II log-power, and WIND/WAVES type II descriptors (inverse delay after flare onset, log-power, and a combined type II/IV score X_W) describe shock formation in the corona and interplanetary space [7, 22]. A WIND burst is temporally matched to the flare with a quality $q_W \in \{0, \frac{1}{2}, 1\}$ (none, weak, strong); for unmatched bursts ($q_W = 0$) the WIND amplitudes are set to zero. To suppress interplanetary bursts that originate in another region, we require a contemporaneous microwave burst: with $H_{\mathrm{RSTN}} = 1$ if a burst is detected at any RSTN frequency $\geq$1415 MHz (see below),

$$B_{\mathrm{RSTN/WIND}} = \ln\Bigl(1 + \sum_{\nu \geq 1.4\,\mathrm{GHz}} I_\nu\Bigr)\, q_W\, H_{\mathrm{RSTN}},$$
$$R_{\mathrm{coh}} = \begin{cases} 0, & \text{weak WIND match and } H_{\mathrm{RSTN}} = 0, \\ X_W + 0.25 \ln\bigl(1 + \sum_{\nu \geq 1.4\,\mathrm{GHz}} I_\nu\bigr), & \text{otherwise.} \end{cases} \tag{5}$$

RSTN microwaves. Microwave gyrosynchrotron emission at $\nu \gtrsim 1$ GHz traces flare-accelerated electrons, whereas the decimetric channels (245–610 MHz) are dominated by plasma emission [12]. For every frequency we subtract the pre-flare background $S_{\nu,0}$, the median flux density in $[t_0 - 30\,\mathrm{min}, t_0)$, and compute

$$S_\nu = \max_{t_0 \leq t \leq t_1 + 10\,\mathrm{min}} \bigl[S_\nu(t) - S_{\nu,0}\bigr], \qquad I_\nu = \int_{t_0}^{t_1 + 10\,\mathrm{min}} \max\bigl[S_\nu(t) - S_{\nu,0}, 0\bigr]\, dt, \tag{6}$$

in sfu and sfu min (1 sfu $= 10^{-22}\,\mathrm{W\,m^{-2}\,Hz^{-1}}$). A burst is "detected" if $S_\nu > \max(5\sigma_0, 10\,\mathrm{sfu})$, where σ_0 is the robust (MAD-based) background scatter. The one-minute data are averages over the available stations (Sagamore Hill, San Vito, Learmonth, Palehua). The model uses $\log_{10}(1 + S_\nu)$ at the eight frequencies and the logarithms of the integrated high-frequency ($\geq$1415 MHz) and decimetric ($\leq$610 MHz) fluences. RSTN data exist for 919 flares.

Active region. The NOAA SRS of the flare day gives the sunspot area A_{AR} (millionths of the hemisphere, MSH), the magnetic (Mount Wilson) class, encoded as $M_{\mathrm{AR}} = 1, 2, 3, 4, 5$ for α, β, $\beta\gamma$, $\beta\delta$, $\beta\gamma\delta$, and the δ-spot flag Δ_δ (757 flares matched by NOAA region number). We also use

$$E_{\mathrm{AR}} = A_{\mathrm{AR}} \ln(1 + G), \qquad C_{\mathrm{AR}} = (M_{\mathrm{AR}} + 2\Delta_\delta + 0.01 A_{\mathrm{AR}}) \ln(1 + G), \tag{7}$$

with $G = F_{\mathrm{peak}}/10^{-8}\,\mathrm{W\,m^{-2}}$, and the flare history H_{24}, the sum of G for all C1.0+ flares of the same region in the preceding 24 h.

These predictors describe the eruption within $\approx$1–2 h of the flare peak. The forecast thus has the character of a short-term warning issued after the CME is observed; for 7 of the 94 SEP flares the proton onset precedes the reference time of the associated CME, which is a limitation of the post-eruption setting.

4 Classification and verification methods

4.1 Classifiers

We compare three probabilistic classifiers. *CatBoost* [25] is a gradient-boosting ensemble of oblivious decision trees that builds an additive log-odds model $F_M(\mathbf{x}) = F_0 + \eta \sum_{m=1}^{M} f_m(\mathbf{x})$, $p = \sigma(F_M)$, by Newton steps on the weighted logarithmic loss

$$\mathcal{L} = -\sum_i w_i \left[y_i \ln p_i + (1 - y_i) \ln(1 - p_i) \right], \qquad w_i = \begin{cases} N_0/N_1, & y_i = 1, \\ 1, & y_i = 0, \end{cases} \tag{8}$$

where N_1 and N_0 are the numbers of SEP and non-SEP flares in the training set. For a leaf R_j the value is $\gamma_j = -\sum_{i \in R_j} g_i / (\sum_{i \in R_j} h_i + \lambda)$ with $g_i = w_i(p_i - y_i)$, $h_i = w_i p_i (1 - p_i)$ and L_2 regularisation λ. We use $M = 800$ trees of depth 3, $\eta = 0.03$, $\lambda = 30$ and random subspaces of 50% of the predictors; missing values are handled natively. The *Random Forest* [26] averages 600 Gini-split trees (maximum depth 8, at least five flares per leaf, $\sqrt{d}$ predictors per split) grown on bootstrap samples with balanced class weights. The physical *logistic baseline* is a class-weighted logistic regression on six standardised predictors: $\log_{10} F_{\mathrm{peak}}$, $\log_{10} \Phi_{\mathrm{SXR}}$, v_{CME}, w_{CME}, C_{W45} and the ground type II flag. For the Random Forest and the logistic model, missing values are replaced by training-set medians and flagged by indicator variables. The hyper-parameters were fixed in advance and were not tuned on the evaluation data.

Class weighting raises the predicted probabilities. Since weighting the positives by $w = N_0/N_1$ multiplies the odds by w, calibrated probabilities are recovered by the prior correction [27]

$$p^* = \frac{p/w}{p/w + 1 - p}. \tag{9}$$

All probabilities quoted below are corrected in this way. The correction is monotonic and does not affect the ranking scores.

4.2 Verification scores

Ranking quality is measured by the area under the receiver operating characteristic (ROC AUC) and by the average precision (AP, the area under the precision–recall curve), which is more informative for rare events [28]; the climatological value of AP equals the event rate, 0.098. Probabilistic accuracy is measured by the Brier skill score $\mathrm{BSS} = 1 - \mathrm{BS}/\mathrm{BS}_{\mathrm{clim}}$. For a binary alert issued when $p^* \geq \tau$, with the contingency counts TP, FP, FN, TN,

$$\mathrm{POD} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}, \quad \mathrm{FAR} = \frac{\mathrm{FP}}{\mathrm{TP} + \mathrm{FP}}, \quad \mathrm{TSS} = \mathrm{POD} - \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}},$$
$$\mathrm{HSS} = \frac{2(\mathrm{TP} \cdot \mathrm{TN} - \mathrm{FP} \cdot \mathrm{FN})}{(\mathrm{TP} + \mathrm{FN})(\mathrm{FN} + \mathrm{TN}) + (\mathrm{TP} + \mathrm{FP})(\mathrm{FP} + \mathrm{TN})}. \tag{10}$$

The true skill statistic (TSS) [29] does not depend on the event rate and is recommended for comparing forecasts across samples [30]; the Heidke skill score (HSS) and FAR depend on it.

4.3 Resampling design

Repeated nested cross-validation. The sample is divided into five stratified folds; each model is trained on four folds and predicts the fifth, which gives an out-of-fold probability for every flare. The alert threshold τ is selected *inside* the four training folds by an inner five-fold cross-validation, maximising either TSS or HSS; it is then applied unchanged to the held-out fold. The procedure is repeated ten times with different random partitions. We report means and standard deviations over the repeats and, for the first repeat, 95% bootstrap confidence intervals (2000 resamples of flares).

Chronological test. To check the stability across solar cycles, the models were trained on the flares of 1997–2010 (569 flares, 63 SEP; mostly cycle 23) and tested on 2011–2022 (395 flares, 31 SEP; cycles 24–25). The threshold is selected by cross-validation within the training period.

Physical attribution. The contribution of each predictor group is measured by ablation (retraining without the group, or with that group only; five repeats of five-fold cross-validation) and by Shapley additive explanations (SHAP) [31] computed with the exact tree algorithm of CatBoost for models trained on the full sample. SHAP values are additive contributions to the log-odds of the individual forecast.

5 Results

5.1 Distribution of the events

Figure 1 shows the two physical quantities that separate the classes most clearly. SEP-productive flares are concentrated at western longitudes (median W27, interquartile range E02–W59, against W01 and E44–W47 for the other flares), in agreement with the Parker-spiral connection, although 22 of the 84 SEP flares with known position (26%) lie in the eastern hemisphere, where only the strongest and widest shocks can reach the Earth-connected field lines. The associated CMEs are faster and wider: the median speed is $1640\,\mathrm{km\,s^{-1}}$ ($900\,\mathrm{km\,s^{-1}}$ for non-SEP flares) and the median width is 360° (46°). The flare peak flux matters less: the medians differ by a factor of ≈2.5 and the classes overlap broadly (Fig. 1b), whereas the SXR fluence differs by a factor of ≈5.6. The probability of an SEP event is 28% for the 163 X-class and 6% for the M-class flares of the sample. Table 1 lists the coverage and the class medians of all predictors.

Table 1 Predictors used by the classifiers: coverage (percentage of the 964 flares with a measured value) and medians for non-SEP and SEP-productive flares

Predictor	Cov., %	Median, no SEP	Median, SEP
Flare SXR			
$\log_{10} F_{\mathrm{peak}}$ (W m^{-2})	100	−4.44	−4.03
$\log_{10} \Phi_{\mathrm{SXR}}$ (J m^{-2})	100	−1.55	−0.796
$t_{\mathrm{p}} - t_0$ (min)	100	11	24
$t_1 - t_0$ (min)	100	20	45
$R_{\mathrm{XS},10}$	63	0.163	0.167
D_{XS} (min^{-1})	63	0.0137	0.0132
D_{SXR} (min^{-1})	63	0.0784	0.0546

Predictor	Cov., %	Median, no SEP	Median, SEP
Longitude/connection			
λ (deg)	88	1	27
$C_{\mathrm{W45}} = \cos(\lambda - 45^\circ)$	88	0.701	0.866
CME kinematics			
v_{CME} (km s^{-1})	99	904	1644
w_{CME} (deg)	99	46	360
S_{CME}	99	25.7	42.5
Type II/IV radio			
type II flag (ground)	100	0	1
type II log power	100	0	1.9
type IV flag	100	0	0
$1/\Delta t_{\mathrm{II}}^{\mathrm{WIND}}$ (clean)	100	0	0.0317
WIND type II log power (clean)	100	0	1.82
$X_{W,\mathrm{comb}}$ (clean)	100	0	2.1
$B_{\mathrm{RSTN/WIND}}$	100	0	6.28
R_{coh}	100	2.08	4.54
RSTN microwaves			
$\log_{10}(1 + S_{245})$	95	1.58	2.82
$\log_{10}(1 + S_{410})$	93	1.44	2.95
$\log_{10}(1 + S_{610})$	93	1.55	2.9
$\log_{10}(1 + S_{1415})$	95	1.61	2.75
$\log_{10}(1 + S_{2695})$	94	1.88	2.96
$\log_{10}(1 + S_{4995})$	93	2.05	3.1
$\log_{10}(1 + S_{8800})$	92	2.13	3.11
$\log_{10}(1 + S_{15400})$	91	2.03	2.88
$\log_{10}(1 + \sum_{\nu \geq 1.4\,\mathrm{GHz}} I_\nu)$	95	3.59	4.81
$\log_{10}(1 + \sum_{\nu \leq 0.6\,\mathrm{GHz}} I_\nu)$	95	3.05	4.39
Active region			
A_{AR} (MSH)	79	375	500
M_{AR}	79	3	3
Δ_δ	79	0	0
E_{AR}	79	3266	4640
C_{AR}	79	72.7	94.2
H_{24}	81	2590	1370

5.2 Cross-validated skill

Table 2 and Fig. 2 summarise the ten repetitions of nested cross-validation. All three classifiers rank the eruptions well: ROC AUC is 0.935 ± 0.004 for CatBoost, 0.932 ± 0.003 for the Random Forest and 0.924 ± 0.002 for the six-predictor logistic baseline. The average precision, which is more sensitive to the ordering of the rare positives, is 0.61, 0.60 and 0.56, respectively, against the climatological value of 0.098. The prior-corrected CatBoost and logistic probabilities have the same Brier skill score (BSS = 0.33); the Random Forest probabilities are less sharp (BSS = 0.24). The reliability diagrams (Fig. 2c) show that the corrected probabilities are approximately

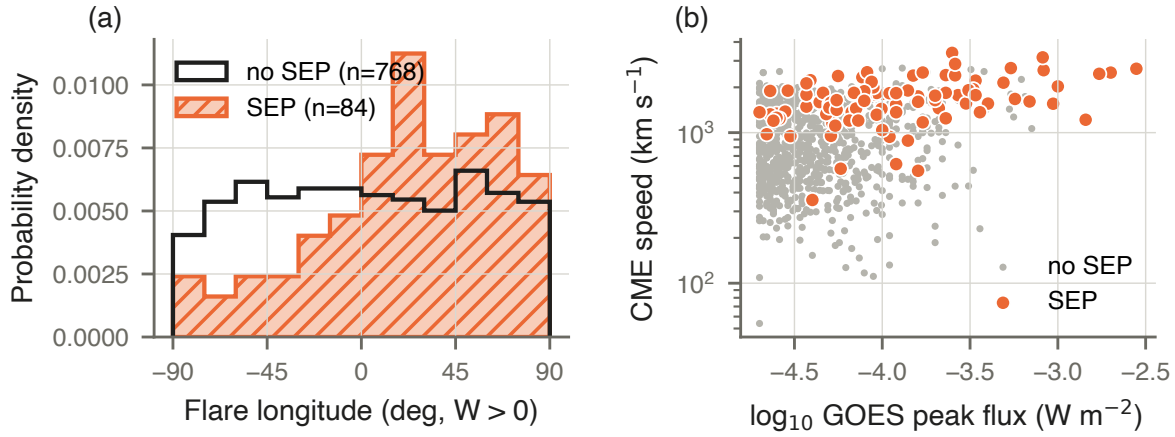


Fig. 1 (a) Distribution of the heliographic longitude of SEP-productive flares (hatched) and of the other flares (solid line). (b) Speed of the associated CME against the GOES peak flux; large filled circles mark SEP-productive flares

Table 2 Repeated nested cross-validation (10 repeats of 5 folds; mean ± standard deviation over the repeats). The threshold is selected in the inner cross-validation by maximising TSS or HSS. TP/FP: mean numbers of warned SEP flares and false alarms (94 SEP flares in total)

Model	Threshold	ROC AUC	AP	BSS	POD	FAR	TSS	HSS	TP/FP
CatBoost	max HSS	0.935±0.004	0.612±0.022	0.33±0.02	0.72±0.03	0.51±0.02	0.64±0.03	0.53±0.02	68/71
Logistic baseline	max HSS	0.924±0.002	0.555±0.006	0.33±0.00	0.61±0.03	0.53±0.03	0.54±0.02	0.47±0.02	57/66
Random Forest	max HSS	0.932±0.003	0.602±0.018	0.24±0.01	0.72±0.06	0.52±0.01	0.64±0.05	0.52±0.02	68/73
CatBoost	max TSS	0.935±0.004	0.612±0.022	0.33±0.02	0.89±0.03	0.62±0.02	0.74±0.02	0.46±0.02	84/136
Logistic baseline	max TSS	0.924±0.002	0.555±0.006	0.33±0.00	0.94±0.01	0.65±0.01	0.75±0.01	0.43±0.01	88/161
Random Forest	max TSS	0.932±0.003	0.602±0.018	0.24±0.01	0.89±0.01	0.60±0.01	0.74±0.01	0.48±0.01	84/128

calibrated below $p^* \approx 0.5$; the few forecasts with larger probabilities are slightly over-confident.

The choice of the alert threshold defines the operating point (Fig. 3). Maximising TSS in the inner cross-validation gives a low threshold ($p^* \approx 0.03$ for CatBoost) and a high-sensitivity regime: POD $= 0.90 \pm 0.03$, FAR $= 0.62 \pm 0.02$, TSS $= 0.74 \pm 0.02$. Maximising HSS gives $p^* \approx 0.13$ and a balanced regime: POD $= 0.72 \pm 0.03$, FAR $= 0.51 \pm 0.02$, TSS $= 0.64 \pm 0.03$, HSS $= 0.53 \pm 0.02$; on average 68 of the 94 SEP flares are warned with 71 false alarms. The bootstrap 95% intervals for this regime (first repeat) are POD 0.71–0.87, FAR 0.45–0.61 and HSS 0.45–0.61. The gradient-boosting and Random-Forest models are statistically indistinguishable. Both improve on the logistic baseline mainly in the balanced regime (HSS 0.52–0.53 against 0.47, POD 0.72 against 0.61 at similar FAR) and in AP; in the high-sensitivity regime the baseline is as good (TSS = 0.75). The bulk of the skill therefore comes from a few physical quantities – SXR fluence and peak flux, CME speed and width, magnetic connection and the type II burst – and the ensembles add information mostly from the microwave and active-region predictors, which matters when false alarms must be limited.

5.3 Chronological test

The models trained on 1997–2010 retain their skill on the independent interval 2011–2022 (Table 3). ROC AUC is 0.943 for CatBoost, 0.948 for the Random Forest and 0.938 for the logistic

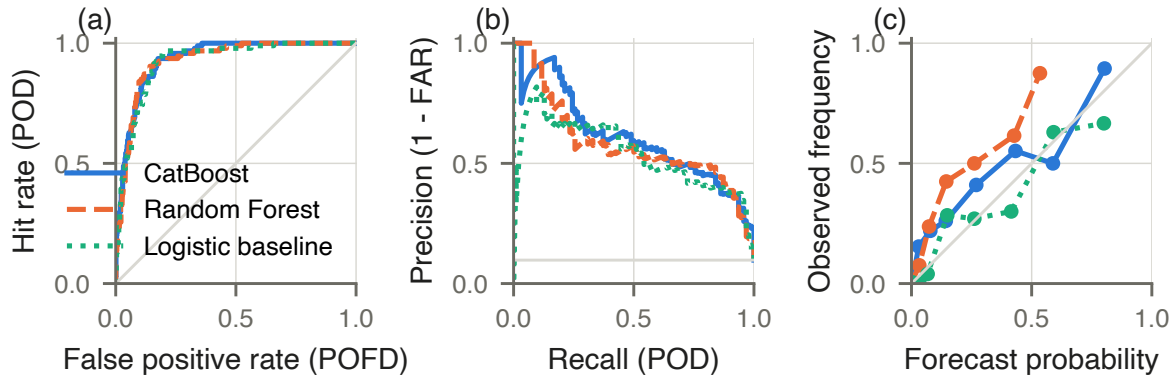


Fig. 2 Out-of-fold verification of the three classifiers (first cross-validation repeat): (a) ROC curves; (b) precision–recall curves, the horizontal line is the climatological event rate; (c) reliability diagram of the prior-corrected probabilities. Solid line: CatBoost; dashed: Random Forest; dotted: logistic baseline

Table 3 Chronological test: training on 1997–2010 (569 flares, 63 SEP), test on 2011–2022 (395 flares, 31 SEP). Brackets give 95% bootstrap intervals

Model	Threshold	ROC AUC	AP	TP/FP/FN	POD	FAR	TSS	HSS
CatBoost	max HSS	0.943	0.53	25/30/6	0.81 [0.66, 0.93]	0.55 [0.41, 0.67]	0.72 [0.57, 0.85]	0.53
Logistic baseline	max HSS	0.938	0.56	12/6/19	0.39 [0.23, 0.57]	0.33 [0.12, 0.57]	0.37 [0.21, 0.55]	0.46
Random Forest	max HSS	0.948	0.65	24/30/7	0.77 [0.61, 0.91]	0.56 [0.42, 0.69]	0.69 [0.53, 0.83]	0.52
CatBoost	max TSS	0.943	0.53	27/40/4	0.87 [0.74, 0.97]	0.60 [0.48, 0.71]	0.76 [0.62, 0.87]	0.50
Logistic baseline	max TSS	0.938	0.56	28/43/3	0.90 [0.78, 1.00]	0.61 [0.49, 0.72]	0.79 [0.66, 0.88]	0.49
Random Forest	max TSS	0.948	0.65	28/42/3	0.90 [0.78, 1.00]	0.60 [0.49, 0.71]	0.79 [0.66, 0.88]	0.50

baseline. With the thresholds selected within the training period, CatBoost warns 25 of the 31 test events with 30 false alarms (POD = 0.81, FAR = 0.55, TSS = 0.72, HSS = 0.53) in the HSS-optimal regime, and 27 events with 40 false alarms in the TSS-optimal regime. These values are consistent with the cross-validated ones within the bootstrap uncertainties, which are large because the test set contains only 31 events. The event rate of the sample is lower in cycles 24–25 (7.8%) than in cycle 23 (11.1%), in line with the weaker activity of cycle 24; the TSS, which does not depend on the event rate, is the appropriate score for this comparison. The HSS-optimal threshold of the logistic baseline does not transfer (POD drops to 0.39), whereas the tree ensembles, whose probabilities depend less on extrapolation of a linear predictor, keep stable operating points.

5.4 Physical attribution

The ablation experiment (Table 4, Fig. 4a) separates the predictor groups by the information they carry alone and by the information that the other groups cannot replace. The sweep-frequency radio group (ground and WIND/WAVES type II and IV) is the strongest single group (ROC AUC = 0.91, AP = 0.47) and the only one whose removal clearly degrades the forecast (AUC from 0.935 to 0.913, AP from 0.62 to 0.54, BSS from 0.34 to 0.25). This agrees with the close association between decametric–hectometric type II bursts and large SEP events [7, 22]: a type II burst signals that the CME drives a shock at heights where protons are efficiently accelerated and released. The SXR, microwave and CME groups are each informative alone

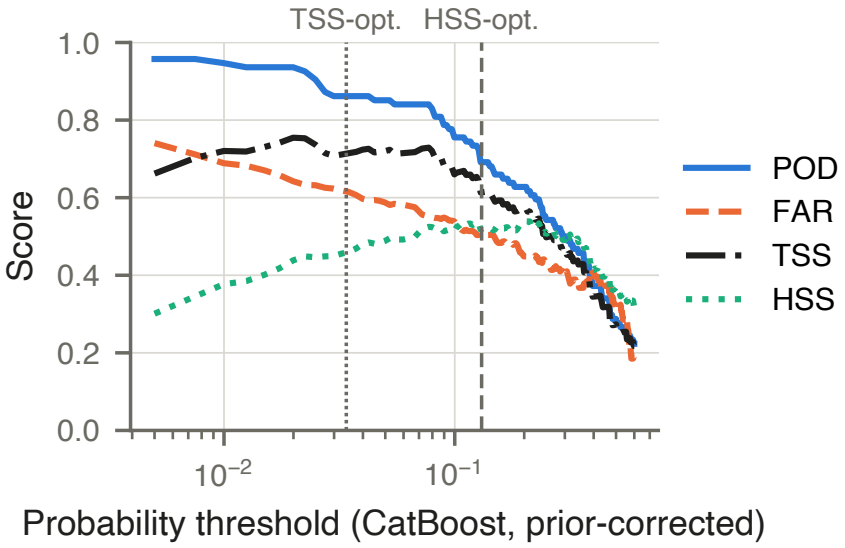


Fig. 3 POD, FAR, TSS and HSS of the out-of-fold CatBoost forecasts as functions of the probability threshold. The vertical lines mark the mean thresholds selected by the inner cross-validation for maximum TSS and maximum HSS

(AUC = 0.82–0.86) but largely redundant with each other and with the radio group: removing any one of them changes AUC by less than 0.01. This is the statistical expression of the "big flare syndrome" [5]: energetic eruptions are large in all observables at once.

The longitude group behaves differently. Alone it has almost no skill (AUC = 0.55), because most well-connected flares do not produce SEP events, yet removing it lowers AP from 0.62 to 0.54 – as much as removing the radio group. The magnetic connection is thus not a predictor of the event by itself but a condition that modulates the effect of the shock strength, which is exactly the role assigned to it by the Parker-spiral picture. The SHAP dependence (Fig. 5a) has the form of a step: the contribution to the log-odds changes from ≈ -0.4 east of E20 to $\approx +0.2$ west of the central meridian and stays nearly constant up to the west limb; the model does not find a narrow optimum near W45–W60, which suggests that the broad longitudinal extent of CME-driven shocks matters more than the precise footpoint [6, 2]. The active-region descriptors from the SRS add little (AUC = 0.64 alone, −0.001 when removed), which is not surprising, as the sunspot area and the magnetic class describe the energy reservoir and not the eruption itself. Restricting the CME association to the interval −60 to +180 min around the flare peak does not change the skill (AUC = 0.935).

The SHAP dependences of the other leading predictors (Fig. 5b–d) are monotonic and have characteristic transition values that agree with physical expectations. The contribution of the CME speed changes sign at $v_{\mathrm{CME}} \approx 1000\,\mathrm{km\,s^{-1}}$ and keeps increasing up to $\approx 2500\,\mathrm{km\,s^{-1}}$, consistent with the requirement of a super-Alfvénic shock in the low corona, where the Alfvén speed is several hundred to $\sim$1000 $\mathrm{km\,s^{-1}}$ [7, 24]. The SXR fluence contribution changes sign at $\Phi_{\mathrm{SXR}} \approx 0.05\,\mathrm{J\,m^{-2}}$, i.e. for an M5 flare lasting ≈15 min or a long-duration M2 event; long-duration SXR events are known to be associated with large CMEs and SEP events [8, 5]. The WIND/WAVES contribution saturates once a type II burst is matched to the flare ($X_W \gtrsim 1$), so the presence of the burst matters more than its intensity. The ranking by mean |SHAP|

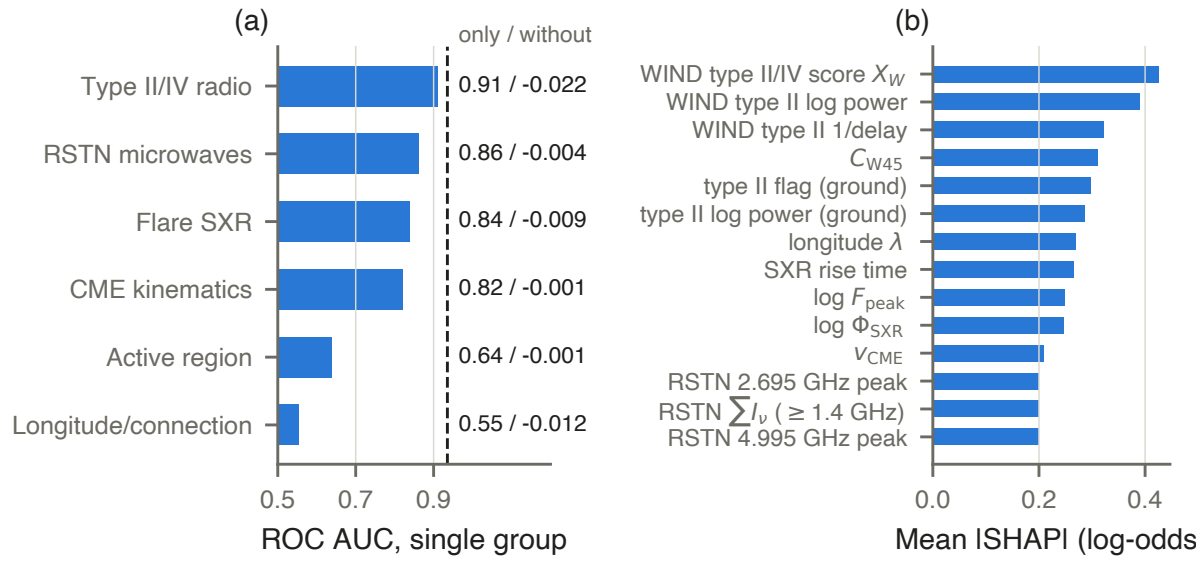


Fig. 4 (a) ROC AUC of CatBoost trained on a single predictor group (bars); the numbers give this value and the change of ROC AUC when the group is removed from the full model, whose AUC is shown by the dashed line. (b) Mean absolute SHAP value of the 14 most important predictors (CatBoost, full sample)

(Fig. 4b) is led by the WIND/WAVES type II descriptors, the connection proxy, the ground type II burst, the longitude and the SXR rise time, peak flux and fluence, followed by the CME speed and the microwave peak flux densities at 2.7 and 5 GHz.

5.5 Missed events and false alarms

The detection efficiency increases with the intensity of the proton event (Fig. 6a): in the HSS-optimal regime CatBoost warns 59% of the events with peak flux 10–30 pfu and 85% of those above 1000 pfu. It also increases toward the west (Fig. 6b), from ≈60–70% for eastern and central sources to ≈80–85% at W30–W90. Weak events from poorly connected sources are therefore the main source of misses, as expected physically: their flux at the Earth is close to the 10 pfu threshold, and small changes of the transport conditions decide whether the threshold is crossed. Several missed events are associated in the initial table with narrow CMEs ($w_{\mathrm{CME}} < 30°$), although the eruption produced a halo CME (e.g. the X2.0 flare of 2000 November 24). These cases reflect the CME association rule (the fastest CME in a 10-h window), which sometimes selects a different, unrelated CME, and indicate that a position-angle consistency check would further improve the forecasts.

The false alarms are physically similar to the SEP-productive eruptions. Of the 72 non-SEP flares that CatBoost alerted in at least half of the repeats (HSS-optimal regime), half are X-class flares, two thirds are accompanied by halo CMEs, and the median CME speed is 1560 km s^{-1}. To test whether some of them are sub-threshold events, we compared the flares of 2010–2022 with the proton-event catalogue of the Skobeltsyn Institute of Nuclear Physics of Moscow State University (SINP MSU) [32] and its continuation for cycle 25, which lists >10 MeV enhancements down to ≈1 pfu. Of the 25 persistent false alarms of that period, 9 (36%) were followed by a catalogued proton enhancement, compared with 24 of the 341 other non-SEP flares (7%); two

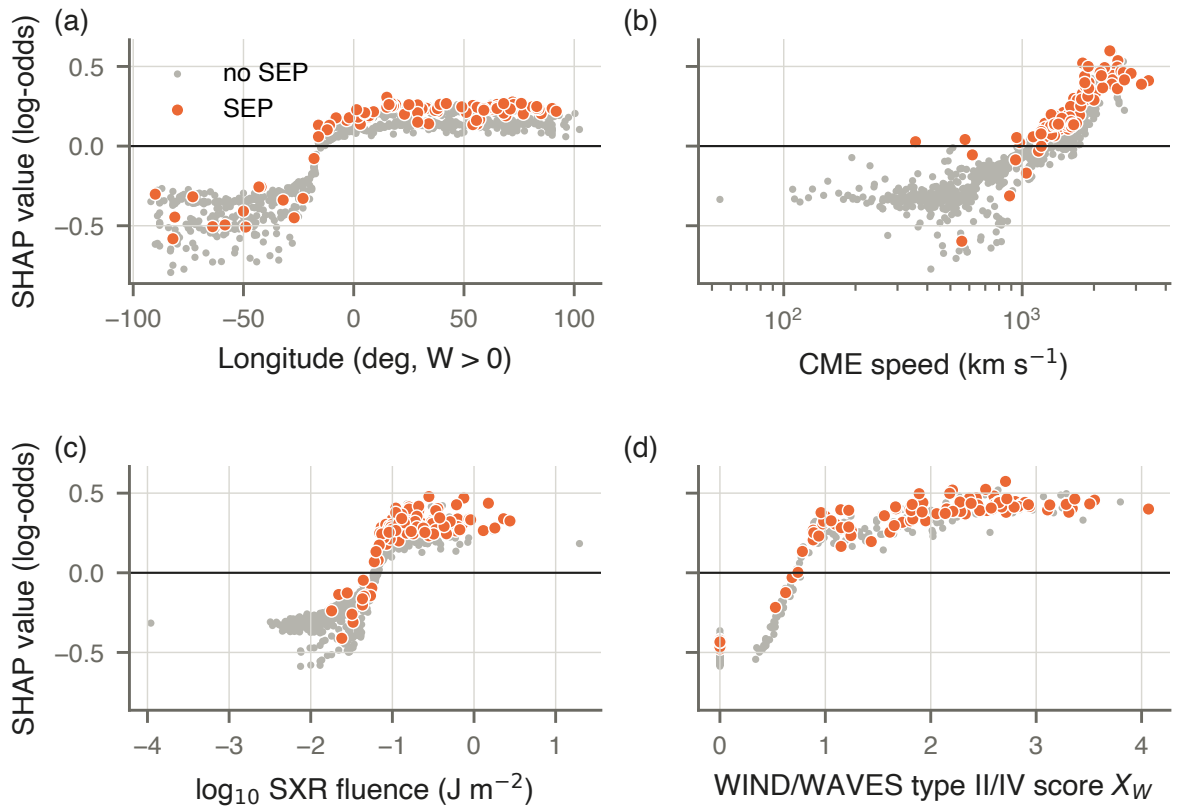


Fig. 5 SHAP dependence plots: contribution of a single predictor to the log-odds of the CatBoost forecast against its value, for (a) flare longitude, (b) CME speed, (c) SXR fluence and (d) the WIND/WAVES type II/IV score. Small grey points: non-SEP flares; large filled circles: SEP-productive flares

of them (2022 April 2 and 2022 August 27, 32 and 27 pfu) are SEP events in which the SINP catalogue and SWPC name different parent flares. A substantial fraction of the false alarms is thus produced by eruptions that did accelerate protons, but below the 10 pfu threshold or with an ambiguous parent flare. The FAR values reported here are therefore an upper limit to the rate of physically wrong forecasts.

5.6 Effect of the data treatment on the reported skill

To quantify the effect of the corrections of Sect. 2.3, the CatBoost model of the preliminary version of this study (25 predictors computed from the initial table, the same hyper-parameters) was re-evaluated in four configurations (Table 5; five repeats of five-fold cross-validation each). Configuration A reproduces the original protocol: initial table with duplicate records, heuristic labels, and a threshold chosen on the out-of-fold predictions themselves to maximise POD at FAR ≤ 0.67. It gives ROC AUC $= 0.83$ and POD $= 0.58$ at FAR $= 0.66$. When the threshold is selected by nested cross-validation (B), the ranking is unchanged but the contingency scores become those that a user would actually obtain; with the TSS criterion, FAR rises to 0.79. Removing the duplicate flare records (C) raises ROC AUC to 0.85, and replacing the heuristic labels by the SWPC parent-flare associations (D) raises it to 0.93 and AP from 0.39 to 0.61. The recomputed predictors of Sect. 3 add a further small gain (Table 2). The largest single effect is therefore the identification of the parent flare: a heuristic label attached to the strongest flare

Table 4 Ablation of predictor groups (CatBoost; 5 repeats of 5-fold cross-validation, mean ± standard deviation)

Predictor set	N	ROC AUC	AP	BSS
all features	36	0.935±0.003	0.621±0.010	0.34±0.01
without Flare SXR	29	0.927±0.004	0.586±0.014	0.31±0.02
without Longitude/connection	34	0.924±0.005	0.543±0.021	0.28±0.02
without CME kinematics	33	0.934±0.005	0.602±0.017	0.33±0.01
without Type II/IV radio	28	0.913±0.004	0.540±0.018	0.25±0.02
without RSTN microwaves	26	0.931±0.003	0.626±0.009	0.35±0.01
without Active region	30	0.934±0.004	0.591±0.021	0.32±0.01
only Flare SXR	7	0.839±0.006	0.341±0.014	0.12±0.01
only Longitude/connection	2	0.553±0.014	0.114±0.003	-0.02±0.01
only CME kinematics	3	0.821±0.007	0.369±0.017	0.17±0.01
only Type II/IV radio	8	0.910±0.008	0.472±0.022	0.27±0.02
only RSTN microwaves	10	0.863±0.007	0.355±0.023	0.16±0.02
only Active region	6	0.640±0.016	0.188±0.012	0.02±0.01
logistic baseline set	6	0.917±0.004	0.509±0.023	0.27±0.02
all features, strict CME timing	36	0.935±0.005	0.612±0.011	0.33±0.02

Table 5 Effect of the data treatment on the skill of the CatBoost configuration of the preliminary version (25 predictors). A: initial table, heuristic labels, threshold tuned on the evaluation data (max POD at FAR $\leq$ 0.67); B: as A with nested threshold (max TSS); C: duplicates removed; D: as C with SWPC parent-flare labels. Five repeats of 5-fold cross-validation

Conf.	Flares/SEP	ROC AUC	AP	POD	FAR	TSS	HSS
A	1150/82	0.831±0.007	0.333±0.022	0.58±0.03	0.66±0.01	0.49±0.03	0.37±0.01
B	1150/82	0.831±0.007	0.333±0.022	0.73±0.03	0.79±0.02	0.52±0.02	0.25±0.03
C	964/82	0.852±0.004	0.387±0.019	0.77±0.04	0.73±0.02	0.57±0.03	0.31±0.02
D	964/94	0.928±0.002	0.605±0.014	0.90±0.02	0.61±0.01	0.75±0.02	0.47±0.02

of the preceding day is a noisy proxy of the physical association, and the noise is concentrated in the most active periods, where several large flares occur within one day.

6 Discussion

Comparison with other studies. AbuBakr Ali et al. [16] reported POD = 0.85 and FAR = 0.30 (sweep-frequency radio sample) and POD = 0.76, FAR = 0.31 (fixed-frequency sample) for a Random Forest trained on a balanced sample created by synthetic oversampling. Our best balanced operating point (POD = 0.72, FAR = 0.51) has a comparable detection rate but a higher FAR. The two results are not directly comparable: FAR depends on the event rate, and synthetic oversampling and threshold choices made on the evaluation data both lower the apparent FAR (Sect. 5.6). The rate-independent TSS of our models, 0.64–0.74, and the POD–FAR pairs are in the range reported for operational and research schemes that use flare, CME and radio information [10, 11, 13, 17]. The leading role of the type II burst, CME speed, SXR fluence and longitude found by the SHAP analysis is in agreement with the feature rankings of [16] and [15], and with the physical picture of shock-dominated large SEP events [1, 2].

What machine learning adds. The six-predictor logistic model reaches 99% of the ROC AUC of the tree ensembles and the same TSS in the high-sensitivity regime. The ensembles are useful when false alarms must be limited (HSS 0.53 against 0.47) and for a stable threshold across solar cycles. Their main advantage is the ability to use partially available and redundant ob-

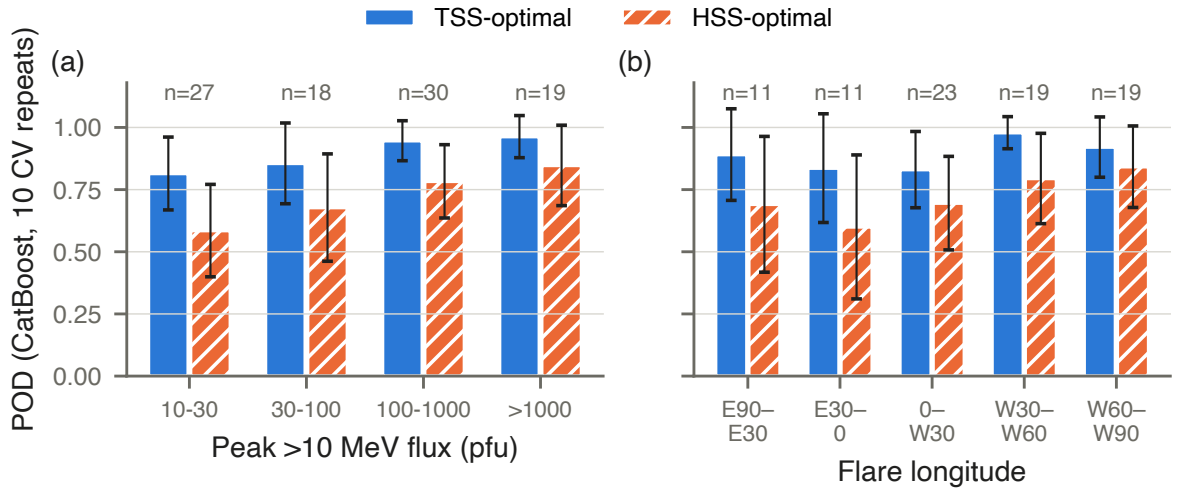


Fig. 6 Probability of detection of CatBoost (averaged over the ten cross-validation repeats) as a function of (a) the peak >10 MeV flux of the proton event and (b) the longitude of the parent flare, for the TSS-optimal (solid) and HSS-optimal (hatched) thresholds. Error bars are binomial 95% intervals; n is the number of events in each bin

servations (microwave spectra, WIND/WAVES parameters, active-region data) without explicit imputation. The physical content of the forecast is, however, concentrated in a few quantities that have been discussed for decades: the shock signature, the CME speed, the energy release and the magnetic connection.

Sensitivity to data treatment. The largest effects on the reported skill are not due to the classifier but to the construction of the sample: the SEP labels, the removal of duplicate flare records and the selection of the alert threshold outside the evaluation data (Table 5). Since these steps are rarely documented in detail, we recommend that SEP classification studies publish the event list with the parent-flare identification and use nested threshold selection.

Limitations. (i) The sample contains only M2.0+ flares with an associated CME and radio data; it is a conditional forecast for strong eruptions and does not cover SEP events from weaker flares, behind-the-limb sources or eruptions without catalogued CMEs. (ii) The CME association (fastest CME in a 10-h window) is time-based; a position-angle check and the use of second-order CME kinematics would reduce association errors (Sect. 5.5). (iii) The WIND/WAVES type II list and the LASCO catalogue are compiled retrospectively; in real time the corresponding information is available with delays of tens of minutes to hours, and for 7 of the 94 SEP flares the proton onset precedes the reference time of the CME. The skill reported here is therefore an upper limit for a real-time implementation. (iv) The binary threshold of 10 pfu treats sub-threshold enhancements as non-events, which inflates the FAR (Sect. 5.5). (v) The test sample of cycles 24–25 contains 31 events, so its scores have bootstrap uncertainties of ±0.1–0.15. Natural extensions are the prediction of the peak proton flux as a continuous quantity, the inclusion of vector-magnetogram (SHARP) parameters for cycles 24–25, and the use of real-time radio and coronagraph data.

7 Conclusions

1. For 964 eruptions of 1997–2022 (M2.0+ flares with CMEs, 94 SEP-productive by the NOAA SWPC parent-flare associations), the probability of a >10 MeV proton event is predicted with ROC AUC = 0.935 ± 0.004, average precision 0.61 and Brier skill score 0.33 by a CatBoost classifier verified with repeated nested cross-validation. The skill is preserved on the independent interval 2011–2022 (ROC AUC = 0.94, TSS = 0.72–0.76).

2. With thresholds selected outside the evaluation data, the forecast warns 72% of the SEP-productive eruptions with FAR = 0.51 (HSS = 0.53), or 90% with FAR = 0.62 (TSS = 0.74). Detection is more efficient for intense events and western sources; about a third of the persistent false alarms in 2010–2022 were followed by proton enhancements below 10 pfu.

3. The forecast information is carried mainly by the radio signature of a coronal and interplanetary shock (type II bursts), the magnetic connection (longitude), the SXR energy release and the CME speed. The longitude acts as a modulating condition, with an effective boundary near E20, and the CME speed contribution changes sign near $1000\,\mathrm{km\,s^{-1}}$. A logistic model on six physical parameters reaches ROC AUC = 0.92; tree ensembles improve mainly the low-false-alarm regime.

4. Duplicate flare records, erroneous flare coordinates, heuristic SEP labels and thresholds tuned on the evaluation data change the apparent skill by an amount comparable to the differences between forecasting methods. The cleaned event list, the code and all results are published to allow independent verification.

Acknowledgements

We thank the NOAA Space Weather Prediction Center and the NOAA National Centers for Environmental Information for the proton-event list, the GOES flare reports, the Solar Region Summaries and the RSTN archive; the operators of the RSTN stations Sagamore Hill, San Vito, Learmonth and Palehua; the CDAW Data Center at NASA Goddard Space Flight Center for the SOHO/LASCO CME catalogue and the WIND/WAVES type II burst list; and the Heliophysics Event Knowledgebase. SOHO is a project of international cooperation between ESA and NASA.

Statements and Declarations

Funding. The authors did not receive support from any organization for the submitted work.

Competing interests. The authors have no relevant financial or non-financial interests to disclose.

Data availability. All data needed to reproduce the results – the initial event table, copies of the public NOAA catalogues used (GOES XRS flare reports, SWPC event reports, both versions of the SWPC proton-event list, Solar Region Summaries, RSTN flux densities around each flare),

the cleaned event sample, all numerical results and the code that reproduces them – are available in the repository `https://github.com/davydenkov/sep-flare-forecast`; the version used in this paper (v1.0.0) is archived at Zenodo, `https://doi.org/10.5281/zenodo.23912985` [35]. The original data are publicly available from NOAA NCEI (`https://www.ngdc.noaa.gov/stp/`), the CDAW Data Center (`https://cdaw.gsfc.nasa.gov/CME_list/`), the NASA SDAC (`https://umbra.nascom.nasa.gov/SEP/`) and the HEK (`https://www.lmsal.com/hek/`).

References

[1] Reames, D. V.: The two sources of solar energetic particles. *Space Sci. Rev.* **175**, 53–92 (2013). https://doi.org/10.1007/s11214-013-9958-9

[2] Desai, M., Giacalone, J.: Large gradual solar energetic particle events. *Living Rev. Sol. Phys.* **13**, 3 (2016). https://doi.org/10.1007/s41116-016-0002-5

[3] Klein, K.-L., Dalla, S.: Acceleration and propagation of solar energetic particles. *Space Sci. Rev.* **212**, 1107–1136 (2017). https://doi.org/10.1007/s11214-017-0382-4

[4] Parker, E. N.: Dynamics of the interplanetary gas and magnetic fields. *Astrophys. J.* **128**, 664–676 (1958). https://doi.org/10.1086/146579

[5] Kahler, S. W.: The role of the big flare syndrome in correlations of solar energetic proton fluxes and associated microwave burst parameters. *J. Geophys. Res.* **87**, 3439–3448 (1982). https://doi.org/10.1029/JA087iA05p03439

[6] Cliver, E. W., Kahler, S. W., Reames, D. V.: Coronal shocks and solar energetic proton events. *Astrophys. J.* **605**, 902–910 (2004). https://doi.org/10.1086/382651

[7] Gopalswamy, N., Yashiro, S., Akiyama, S., et al.: Coronal mass ejections, type II radio bursts, and solar energetic particle events in the SOHO era. *Ann. Geophys.* **26**, 3033–3047 (2008). https://doi.org/10.5194/angeo-26-3033-2008

[8] Belov, A., Garcia, H., Kurt, V., Mavromichalaki, H., Gerontidou, M.: Proton enhancements and their relation to the X-ray flares during the three last solar cycles. *Sol. Phys.* **229**, 135–159 (2005). https://doi.org/10.1007/s11207-005-4721-3

[9] Cliver, E. W.: Flare vs. shock acceleration of high-energy protons in solar energetic particle events. *Astrophys. J.* **832**, 128 (2016). https://doi.org/10.3847/0004-637X/832/2/128

[10] Laurenza, M., Cliver, E. W., Hewitt, J., et al.: 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**, S04008 (2009). https://doi.org/10.1029/2007SW000379

[11] Núñez, M.: Predicting solar energetic proton events ($E > 10$ MeV). *Space Weather* **9**, S07003 (2011). https://doi.org/10.1029/2010SW000640

[12] Zucca, P., Núñez, M., Klein, K.-L.: Exploring the potential of microwave diagnostics in SEP forecasting: the occurrence of SEP events. *J. Space Weather Space Clim.* **7**, A13 (2017). https://doi.org/10.1051/swsc/2017011

[13] Papaioannou, A., Sandberg, I., Anastasiadis, A., et al.: Solar flares, coronal mass ejections and solar energetic particle event characteristics. *J. Space Weather Space Clim.* **6**, A42 (2016). https://doi.org/10.1051/swsc/2016035

[14] Anastasiadis, A., Papaioannou, A., Sandberg, I., et al.: Predicting flares and solar energetic particle events: the FORSPEF tool. *Sol. Phys.* **292**, 134 (2017). https://doi.org/10.1007/s11207-017-1163-7

[15] Lavasa, E., Giannopoulos, G., Papaioannou, A., et al.: Assessing the predictability of solar energetic particles with the use of machine learning techniques. *Sol. Phys.* **296**, 107 (2021). https://doi.org/10.1007/s11207-021-01837-x

[16] AbuBakr Ali, M., Abdelkawy, A. G. A., Shaltout, A. M. K., Beheary, M. M.: Forecasting solar energetic particles using multi-source data from solar flares, CMEs, and radio bursts with machine learning approaches. *Sci. Rep.* **15**, 9546 (2025). https://doi.org/10.1038/s41598-025-92207-1

[17] Whitman, K., Egeland, R., Richardson, I. G., et al.: Review of solar energetic particle prediction models. *Adv. Space Res.* **72**, 5161–5242 (2023). https://doi.org/10.1016/j.asr.2022.08.006

[18] Varma, S., Simon, R.: Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics* **7**, 91 (2006). https://doi.org/10.1186/1471-2105-7-91

[19] Hurlburt, N., Cheung, M., Schrijver, C., et al.: Heliophysics event knowledgebase for the Solar Dynamics Observatory (SDO) and beyond. *Sol. Phys.* **275**, 67–78 (2012). https://doi.org/10.1007/s11207-010-9624-2

[20] Gopalswamy, N., Yashiro, S., Michalek, G., et al.: The SOHO/LASCO CME catalog. *Earth Moon Planets* **104**, 295–313 (2009). https://doi.org/10.1007/s11038-008-9282-7

[21] Bougeret, J.-L., Kaiser, M. L., Kellogg, P. J., et al.: WAVES: the radio and plasma wave investigation on the WIND spacecraft. *Space Sci. Rev.* **71**, 231–263 (1995). https://doi.org/10.1007/BF00751331

[22] Gopalswamy, N., Mäkelä, P., Yashiro, S.: A catalog of type II radio bursts observed by Wind/WAVES and their statistical properties. *Sun Geosph.* **14**, 111–121 (2019). https://doi.org/10.31401/SunGeo.2019.02.03

[23] Struminsky, A. B., Sadovskii, A. M., Grigorieva, I. Yu.: Criteria for forecasting proton events by real-time solar observations. *Geomagn. Aeron.* **64**, 139–149 (2024). https://doi.org/10.1134/S0016793223600984

[24] Kahler, S. W.: The correlation between solar energetic particle peak intensities and speeds of coronal mass ejections: effects of ambient particle intensities and energy spectra. *J. Geophys. Res.* **106**, 20947–20955 (2001). https://doi.org/10.1029/2000JA002231

[25] Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., Gulin, A.: CatBoost: unbiased boosting with categorical features. In: *Advances in Neural Information Processing Systems* **31**, 6638–6648 (2018). arXiv:1706.09516

[26] Breiman, L.: Random forests. *Mach. Learn.* **45**, 5–32 (2001). https://doi.org/10.1023/A:1010933404324

[27] King, G., Zeng, L.: Logistic regression in rare events data. *Polit. Anal.* **9**, 137–163 (2001). https://doi.org/10.1093/oxfordjournals.pan.a004868

[28] Saito, T., Rehmsmeier, M.: The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS ONE* **10**, e0118432 (2015). https://doi.org/10.1371/journal.pone.0118432

[29] Woodcock, F.: The evaluation of yes/no forecasts for scientific and administrative purposes. *Mon. Weather Rev.* **104**, 1209–1214 (1976). https://doi.org/10.1175/1520-0493(1976)104<1209:TEOYFF>2.0.CO;2

[30] Bloomfield, D. S., Higgins, P. A., McAteer, R. T. J., Gallagher, P. T.: Toward reliable benchmarking of solar flare forecasting methods. *Astrophys. J. Lett.* **747**, L41 (2012). https://doi.org/10.1088/2041-8205/747/2/L41

[31] Lundberg, S. M., Erion, G., Chen, H., et al.: From local explanations to global understanding with explainable AI for trees. *Nat. Mach. Intell.* **2**, 56–67 (2020). https://doi.org/10.1038/s42256-019-0138-9

[32] Logachev, Yu. I., Bazilevskaya, G. A., Vlasova, N. A., et al.: Catalog of solar proton events in the 24th cycle of solar activity (2009–2019). Geophysical Center of the Russian Academy of Sciences, Moscow (2022). https://doi.org/10.2205/ESDB-SAD-008

[33] NOAA Space Weather Prediction Center: Solar proton events affecting the Earth environment, 1976–present. NOAA National Centers for Environmental Information (2026). https://www.ngdc.noaa.gov/stp/space-weather/interplanetary-data/solar-proton-events/

[34] NOAA National Centers for Environmental Information: GOES X-ray flare reports, 1975–2017 (2026). https://www.ngdc.noaa.gov/stp/space-weather/solar-data/solar-features/solar-flares/x-rays/goes/xrs/

[35] Davydenkov, I. I., Ozheredov, V. A.: sep-flare-forecast: data and code for solar proton event classification from flare, CME and radio observations (1997–2022), version 1.0.0. Zenodo (2026). https://doi.org/10.5281/zenodo.23912985