The Astrophysical Journal, 974:188 (11pp), 2024 October 20 https://doi.org/10.3847/1538-4357/ad6d57
© 2024. The Author(s). Published by the American Astronomical Society.
OPEN ACCESS

Precise and Accurate Short-term Forecasting of Solar Energetic Particle Events with Multivariate Time-series Classifiers

Sumanth A. Rotti[1], Berkay Aydin[2], and Petrus C. Martens[1]
[1] Department of Physics and Astronomy, Georgia State University, Atlanta, GA 30303, USA; srotti@gsu.edu
[2] Department of Computer Science, Georgia State University, Atlanta, GA 30303, USA
Received 2024 July 8; revised 2024 August 5; accepted 2024 August 7; published 2024 October 11

Abstract

Solar energetic particle (SEP) events are one of the most crucial aspects of space weather that require continuous monitoring and forecasting using robust methods. We demonstrate a proof of concept of using a data-driven supervised classification framework on a multivariate time-series data set covering solar cycles 22, 23, and 24. We implement ensemble modeling that merges the results from three proton channels ($E \geqslant 10$ MeV, 50 MeV, and 100 MeV) and the long-band X-ray flux (1–8 Å) channel from the Geostationary Operational Environmental Satellite missions. Our task is binary classification, such that the aim of the model is to distinguish strong SEP events from nonevents. Here, strong SEP events are those crossing the Space Weather Prediction Center's "S1" threshold of solar radiation storm and proton fluxes below that threshold are weak SEP events. In addition, we consider periods of nonoccurrence of SEPs following a flare with magnitudes $\geqslant$C6.0 to maintain a natural imbalance of sample distribution. In our data set, there are 244 strong SEP events comprising the positive class. There are 189 weak events and 2460 "SEP-quiet" periods for the negative class. We experiment with summary statistic, one-nearest neighbor, and supervised time-series forest (STSF) classifiers and compare their performance to validate our methods for prediction windows from 5 minutes up to 60 minutes. We find the STSF model to perform better under all circumstances. For an optimal classification threshold of $\approx$0.3 and a 60 minutes prediction window, we obtain a true skill statistic TSS = 0.850, Heidke skill score HSS = 0.878, and Gilbert skill score GSS = 0.783.

Unified Astronomy Thesaurus concepts: Solar energetic particles (1491); Space weather (2037)

1. Introduction

The activity of the Sun is considered the prime source of space weather (SWx) that constitutes different eruptive phenomena including solar flares (SFs) and coronal mass ejections (CMEs). A SF is a sudden brightening in the solar atmosphere in a coronal soft X-ray source, which is observed to have close relationships with CMEs (J. Feynman & A. J. Hundhausen 1994) and other observable phenomena such as jets and filament eruptions (C. J. Schrijver & G. L. Siscoe 2010). CMEs are clouds of plasma formed in the lower corona that often move faster (>1000 km s^{-1}) than the ambient solar wind (B. C. Low 1996; P. Chen 2011). Another manifestation of solar activity, constituting the emission of energetic electrons, protons, and heavier ions from the Sun, is called a solar energetic particle (SEP) event (K.-L. Klein & S. Dalla 2017). Typically, large flares and shock fronts of fast CMEs are known to accelerate SEPs and are often considered precursors or parent eruptions (H. V. Cane et al. 1986; S. W. Kahler 1992; N. Gopalswamy et al. 2001, 2008, 2017; C. Marqué et al. 2006; B. Swalwell et al. 2017; E. W. Cliver & E. D'Huys 2018; S. Rotti & P. C. Martens 2023).

The Sun releases ions of tens of MeV energies and more in many SEP events. The motion of such highly charged particles is dictated by magnetic field lines (D. V. Reames 2013). To be geo-effective, SEPs should reach the near-Earth space on a magnetically well-connected path (H. V. Cane et al. 1988). Generally, it is understood that the eruptions at the western side of the Sun have a higher probability of SEPs being geo-effective due to the spiral structure of the interplanetary magnetic field lines, known as the Parker spiral (E. N. Parker 1965; D. V. Reames 1999). Such Earth-bound SEPs are hazardous on many levels. The impacts include severe technological (D. F. Smart & M. A. Shea 1992) and biological effects on various economic scales (C. J. Schrijver & G. L. Siscoe 2010). Although Earth's magnetic field provides us a protective shield from the energetic particles and filters them out from reaching the ground, they can be life-threatening for humans on missions outside of the Earth's magnetosphere and aircraft traveling along polar routes (P. Beck et al. 2005; N. A. Schwadron et al. 2010). For instance, long-lasting strong SEP events pose a radiation hazard to astronauts and electronic equipment in space (P. Jiggens et al. 2019). In addition, the particles getting diverted to Earth's magnetic poles disturb the ionosphere's upper layers, causing disruption to high-frequency radio, GPS, and satellite communications (M. Desai & J. Giacalone 2016).

Over the last four decades, great progress in space exploration has provided near-continuous observations of solar activity from a fleet of advanced space-based satellites (I. G. Usoskin 2017). These observational data should be analyzed in operational contexts to mitigate SWx effects on our human explorers and technological systems (C. H. Jackman & R. D. McPeters 1987). Therefore, we require robust tools and systems to forecast eruptive event occurrences such as SEPs and send warning signals before the event. Researchers across the globe have been focusing on implementing a variety of model-driven techniques for predicting SEP events, mostly concentrating on predicting the characteristics related to peak flux. To forecast SEP event occurrences, physics-based and

Original content from this work may be used under the terms of the Creative Commons Attribution 4.0 licence. Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.

data-driven statistical models have been designed based on the parameters of parent solar eruptions such as SFs and CMEs (M. A. I. Van Hollebeke et al. 1975; S. W. Kahler et al. 2007; A. Posner 2007; C. C. Balch 2008; M. Laurenza et al. 2009; D. Falconer et al. 2011; M. Nunez 2011; M. Dierckxsens et al. 2015; T. Alberti et al. 2017; A. Anastasiadis et al. 2017; A. Papaioannou et al. 2018).

Machine learning (ML) methods have been at the forefront of SEP event forecasting in the last decade due to their success in many other areas of research and operations (S. F. Boubrahimi et al. 2017; A. J. Engell et al. 2017; B. Swalwell et al. 2017; E. Camporeale 2019; S. Aminalragia-Giamini et al. 2021; E. Lavasa et al. 2021; K. Whitman et al. 2022; P. Hosseinzadeh et al. 2024). Our previous work (S. A. Rotti et al. 2024) implemented time-series classifiers on a multivariate time series (MVTS) constituting a solar X-ray and proton fluxes data set of SEP events and showed that an interval-based supervised time-series forest (STSF) model is more efficient in the classification of strong SEP events. Here, one possible area of improvement is to consider an actual class-imbalance scenario, that is, including the phases of non-SEPs and comparing with baseline time-series classification algorithms for the task at hand. Expanding on these ideas, in the present work we implement STSF, summary statistic classifier (SSC) and one-nearest neighbor (1NN) classifiers on an extended data set constituting SEP-quiet periods and analyze their performance. With that, we compare the results and establish a proof of concept of the model to be implemented in a near-real-time (NRT) scenario. That is, we aim to build a high-fidelity (low-risk) and interpretable short-term (low lead time) predictive model for reliable SEP event forecasting systems suitable to operational standards.

SEPs observed in near-Earth space are rare events compared to major flares and CMEs. In addition, the occurrences of SEPs are dependent on precursor solar eruptions (E. W. Cliver et al. 2022; S. Rotti & P. C. Martens 2023). Thus, it makes physical sense to build SEP predictive models based on the parameters of parent solar events. However, in the present work we consider only SFs and decline to use CME data on account of the low quality, such as the projection effects and difficulty of tracking them as three-dimensional objects.

The rest of the paper is organized as follows. Section 2 provides information about our data set and data preparation steps used in this work. Section 3 presents our research methodology, including descriptions of the time-series classification models and data partitioning scheme. Section 4 discusses the training and validation phases of the model and presents the experimental evaluation framework. Lastly, Section 5 provides a summary of our work and future endeavors.

2. Data

In this work, we implement a MVTS data integration, model training and selection pipeline toward building an efficient architecture for short-term SEP event forecasting. We consider a binary classification task between strong SEP and weak or non-SEP events. In the present framework, we consider two classes, "SEP" and "SEP-quiet," where SEP-quiet periods include smaller proton flux increases, as well as the flares that did not lead to SEP events near Earth. With that, we maintain a natural class imbalance between the occurrence and nonoccurrence of strong SEP events. Our methodology for the creation of SEP-quiet samples will be presented in the next section. The strong and weak SEP samples in our work come from the Geostationary Solar Energetic Particle (GSEP) events data set (S. Rotti et al. 2022b), which is discussed later in Section 2.4.

2.1. SEP-quiet Periods

The National Oceanic and Atmospheric Administration (NOAA) has been operating the Geostationary Operational Environmental Satellite (GOES) series in geosynchronous orbits since 1976 (H. H. Sauer 1989; P. L. Bornmann et al. 1996). The space environment monitor (SEM) on board GOES missions measures solar radiation in the X-ray and EUV region and the in situ magnetic field and energetic particle environment (R. N. Grubb 1975). In this work, we utilize the GOES solar X-rays (H. A. Garcia 1994) and energetic proton fluxes (T. Onsager et al. 1996). We create our negative samples using "SEP-quiet" periods in two stages: (i) obtaining lists of SF and SEP events, and (ii) gathering the corresponding GOES flux measurements for each event. We note that we take into account the scaling factors initially implemented by NOAA in the GOES-08 through -15 X-ray measurements. Toward that, we updated our data set by removing the necessary scaling factors as per the NOAA guidelines to get true X-ray fluxes.[3]

NOAA's GOES has traced flaring activity since the beginning of 1974 and offers a catalog with spatial and temporal specifications, flare magnitude, and associated active region (AR) information.[4] However, some flare locations and/or AR numbers are missing from the GOES catalog or probably have known errors (R. O. Milligan & J. Ireland 2018; R. A. Angryk et al. 2020; S. A. Rotti et al. 2020). NOAA categorizes SFs based on soft X-ray peak flux in the wavelength range 1–8 Å. Flare classes from least to most intense are labeled as A, B, C, M, and X, where each category indicates an increase in flare intensity on a logarithmic scale.

2.2. Non-SEP Samples

We select our non-SEP samples such that there are no enhancements or relatively no enhancements in the GOES proton fluxes following flaring episodes. In other words, we identify large flares that do not lead to significant variations in the GOES $\geqslant$10 MeV proton fluxes relative to the background levels. This step helps us increase the sample size over the previous study. At first, we chose M- and X-class flares from the NOAA flare list between 1986 and 2018 that have not been associated with strong SEP events. Then, to add a few more negative samples, we extend our lower limit of a "non-SEP" flaring intensity using a data-driven approach. The GSEP data set consists of carefully identified SF information for 383 associated SEP events. In that, we see a distribution of strengths among 59 SEP events associated with <C6.0 flares, where 38 are weak events and 21 are strong events. Nonetheless, we inspect all of the SEP-associated flares, as shown in Figure 1, and find that <C6.0 flares correspond to the bottom 15th percentile. Hence, we consider C6.0 as our bottom threshold for flares. Then, we add all flares from M1.0 to C6.0 to our earlier flare list and obtain a total of 7981 flares. We filter the preliminary negative sample flares in two steps: (i) we do not consider flares during an ongoing SEP event (either rise or decay), and (ii) we remove all consecutive flares within 11 hr

[3] GOES XRS scaling factors: https://ngdc.noaa.gov/stp/satellite/goes/doc/GOES_XRS_readme.pdf.

[4] The NOAA flare list: https://www.ngdc.noaa.gov/stp/space-weather/solar-data/solar-features/solar-flares/x-rays/.

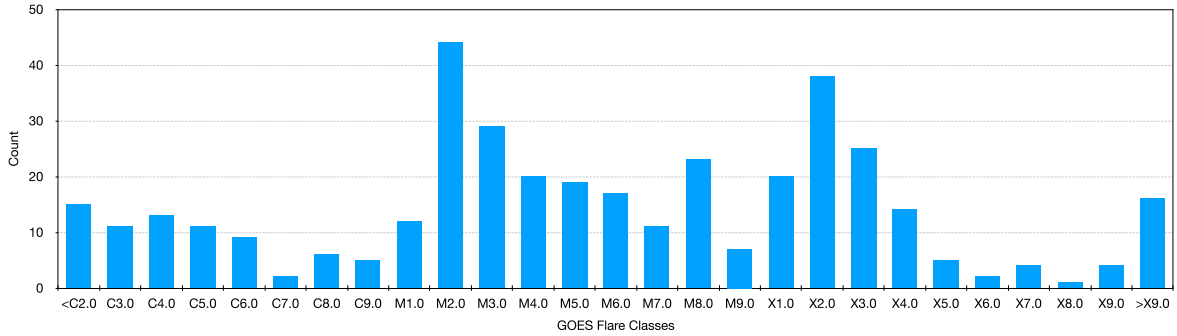

Figure 1. Distribution of soft X-ray flare peak intensities based on GOES flare classification for SEP-associated flares in the GSEP data set. There are a total of 383 SEPs with flare class information, of which 59 events have flare magnitudes less than GOES class C6.0.

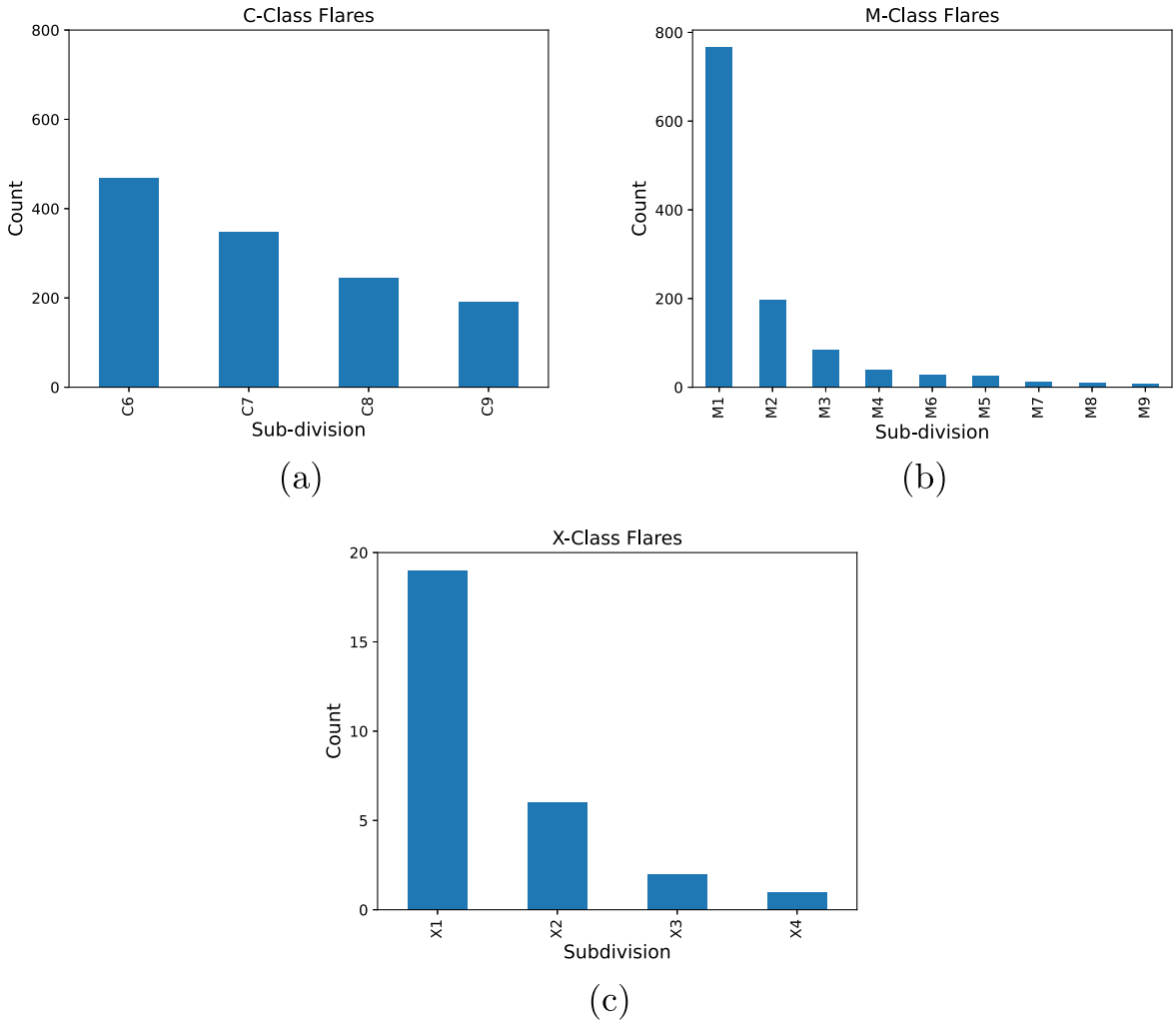

Figure 2. Distribution of the GOES subdivisions of (a) C-class, (b) M-class, and (c) X-class flares present in our "SEP-quiet" periods.

following the onset of the flare at consideration. With this filtering step, our final count of non-SEP flares with an intensity ⩾C6.0 is reduced to 2460. In Figure 2, we show the distribution of the number of SEP-quiet flares in each subdivision of GOES C-class (panel (a)), M-class (panel (b)), and X-class (panel (c)).

K. Whitman et al. (2022) recently organized an SEP model validation (SEPVAL) challenge in 2023.[5] For this, two lists of 33 SEP and 30 non-SEP events from solar cycles 24 and 25 have been provided. In this regard, we cross-verified our non-SEPs with the SEPVAL's list of 15 non-SEP events for solar cycle 24. We do not include any (SEP or non-SEP) events from solar cycle 25 in the present work to make our analysis comparable with earlier studies.

2.3. Time-series Slicing

To generate our data set, we have interpolated and integrated particle and X-ray flux data from several GOES missions. First,

[5] SEPVAL Challenge: https://ccmc.gsfc.nasa.gov/community-workshops/ccmc-sepval-2023/.

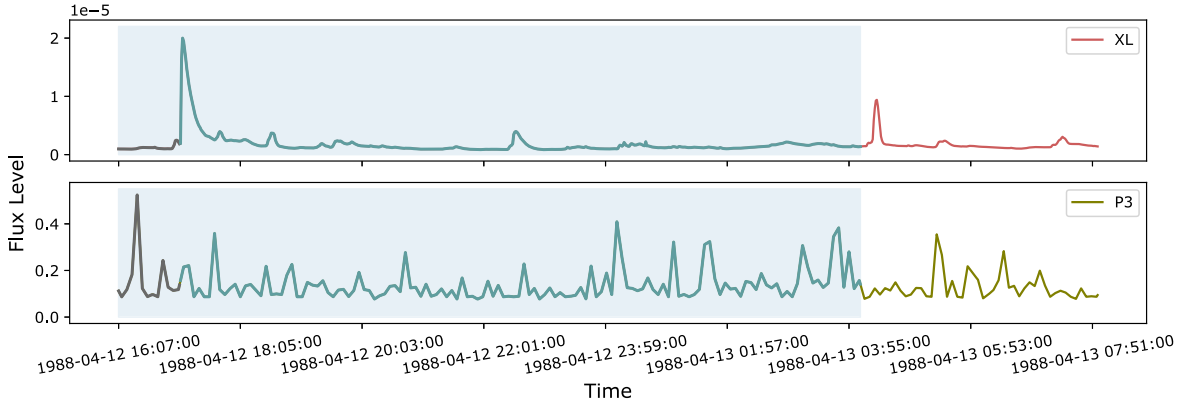

Figure 3. An example of the sampling technique for the non-SEP samples in our data set. We show the XL and P3 fluxes for representation only. The blue shaded region in the plot shows the 12 hr length of the input time series of our sample. Here, the gray shaded line on the left corresponds to the 1 hr of fluxes prior to the (⩾C6.0) flare onset, while the rest of the time series within the blue region covers 11 hr after the onset.

we use the 1 minute averaged X-ray (1–8 Å) fluxes measured by the X-ray sensor (XRS) on board GOES. The archived data are available online from NOAA's website.[6] In addition to X-ray data, we use the following 5 minute averaged integral proton fluxes measured by the SEM suite: (i) $E \geqslant 10$ MeV fluxes corresponding to P3, (ii) $E \geqslant 50$ MeV fluxes corresponding to P5, and (iii) $E \geqslant 100$ MeV fluxes corresponding to P7. We interpolate the proton fluxes to1 minute averages and retain the time stamp of every first measurement. The rationale here is to capture the flare dynamics that occur over a few minutes. In the present work, we follow the same procedure as discussed in S. A. Rotti et al. (2024) to generate our event slices for the additional negative samples, and to ensure that our analysis and models' predictive capabilities are relevant and comparable across different SEP event prediction tasks.

We consider flares as the precursor to the onset of an SEP event. In our earlier statistical study on strong and weak SEP events (S. Rotti & P. C. Martens 2023), we find that 77(±4)% of SEP event onsets occur within 11 hr after the associated flaring eruption. However, sampling flares from the onset will likely lead the classifier to memorize the pattern of finding a spike at the start of the MVTS instance. A simple solution to randomize the sampling is to include 1 hr of data prior to the flare onset because two to three flares may occur within an hour of observation in many instances. With that consideration, the classifiers become free from memorizing to find a spike at the start of the data or at the end of 60 data points corresponding to 1 hr. We accommodate both these aspects while creating our non-SEP time-series sampling because our goal is to develop a highly accurate model predicting strong SEP events. As reported in our previous work (S. A. Rotti et al. 2024), we note that our modeling schema can predict well-connected events in addition to strong proton enhancements. A caveat/drawback of the model would be to classify events that fluctuate around the 10 pfu threshold.

Based on our filtered list of 2460 flares for negative samples, we consider 1 hr of data (all four fluxes mentioned above) before the flare onset and 11 hr afterward for each flaring event (as shown in Figure 3). Therefore, the total length of each MVTS corresponds to 12 hr. As our final data set comes with a cadence of 1 minute, each MVTS is a size-720 sequence of soft X-ray and integral proton fluxes. All time-series data were cleaned for data gaps with linear interpolation and spike removal by replacing what would be considered an outlier (spike) with an interpolated value from moving averages. We note here that the fluxes are predominantly derived from primary GOES missions. Nonetheless, when there are ambiguous measurements in the primary satellite, we consider data from the secondary satellite of the corresponding period of observation. Because we are focusing on short-term predictions, we explore multiple prediction windows: 5, 15, 30, 45, and 60 minutes. In each case, we exclude the respective minutes of observations from the 720-size MVTS at the end by maintaining a constant 11 hr observation window as input to the model. Such a continuous and consistent observation window is needed to train a predictive model that can be established in a NRT operational environment. The prediction window indicates the corresponding time in the future at the end of the observation window for when the model output is valid.

2.4. GSEP Data Set

Our primary source of SEP events comes from the recently published open-source GSEP data set (S. Rotti et al. 2022a), which is available on the Harvard Dataverse. It consists of a carefully vetted SEP events catalog with associated parent eruption metadata and time-series slices. The data set comprises 433 (244 strong and 189 weak) SEP events covering solar cycles 22 to 24. According to the NOAA Space Weather Prediction Center (SWPC), a strong SEP event is defined as having proton intensities $I_P \geqslant 10$ pfu (1 pfu = 1 particle per cm^2 s sr) in the GOES 5 minutes averaged $E \geqslant 10$ MeV integral energy channel for at least three consecutive readings (H. Bain et al. 2021). In the GSEP data set, "weak" SEP events correspond to an enhancement in proton intensities in the $E \geqslant 10$ MeV channel whose peak flux measurements do not reach 10 pfu. The description of the data set and its development can be found in S. Rotti et al. (2022b) and S. Rotti & P. C. Martens (2023).

3. Methodology

We consider the term "SEP events" analogous to solar proton events. While variations exist, positive labels are usually associated with the occurrence of strong SEPs based on the integral proton fluxes recorded by the GOES-P3 channel crossing the 10 pfu threshold. As mentioned earlier, we have

[6] GOES-XRS: https://www.ncei.noaa.gov/data/goes-space-environment-monitor/access/avg/.

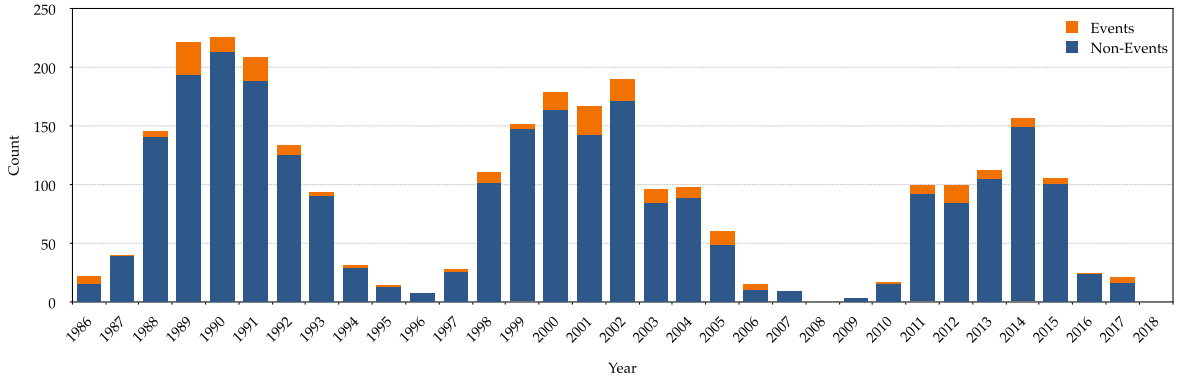

Figure 4. The distribution of SEP and non-SEP samples in our data set between 1986 and 2018. We split our data set into three parts of nonoverlapping years: Partition 1 = 1986–1992, Partition 2 = 1993–2002, and Partition 3 = 2003–2018. Here, each split consists of 998, 974, and 921 samples, respectively. The data set has a cumulative class-imbalance ratio of ≈1:11. In the legend, "Events" are defined as those SEPs crossing the SWPC threshold of 10 pfu in the $E \geqslant 10$ MeV channel, and "Non-Events" consists of both weaker enhancements <10 pfu or no enhancements.

189 weak SEP events in the GSEP data set and we add 2460 non-SEPs to increase the negative samples to 2649. For the positive class, we have 244 strong SEP events. The total size of the data set is 2893 samples with a class-imbalance ratio of ≈1:11. Figure 4 shows the total number of samples in our data set from 1986 to 2018. Here, "Events" are defined as those SEPs crossing the SWPC threshold of 10 pfu in the $E \geqslant 10$ MeV channel, and "Non-Events" constitute both weaker enhancements of <10 pfu and no enhancements. In Section 3.1, we briefly explain the method of time-series classification implemented in the present work.

3.1. Time-series Classification

We examine the performance of time-series classifiers in the framework of a binary classification task for predicting SEP events. Time-series classification uses supervised ML to analyze labeled classes of time-series data and then predicts the class to which a new instance belongs. This is important in SWx predictions, where observational data are analyzed to support NRT operations and classification accuracy becomes critical. Furthermore, for short-term predictions, the algorithm must be highly accurate and robust to be useful. In our previous work (S. A. Rotti et al. 2024), we showed that an ensemble framework of interval-based classifiers achieved high performance in classifying between strong and weak SEP events.

In the present work, we implement three time-series classifiers and compare their performance to establish certainty in our SEP event prediction system. The motive is to see whether the input time-series span leads to a strong SEP event (positive class) or not (negative class). We implement the same column ensemble schema presented in our previous work (S. A. Rotti et al. 2024), where the ensemble estimator allows multiple feature columns of the input to be transformed separately. The statistical features generated by each classifier on samples of the original time series are ensembled to create a single output. Each feature is assigned a score that indicates how informative it is toward predicting the target variable (L. K. Hansen & P. Salamon 1990; R. E. Schapire 1990; M. A. Arbib 2003). The predictions from individual classifiers are then aggregated with equal votes using prediction probabilities. We will give a brief overview of our models in the following sections.

3.2. Summary Statistic Classifier

The feature-based SSC is the first of the two baseline models we have used in this work. This classifier applies a time-series feature transformation to an entire univariate time series using simple summary statistics and builds a random forest using these summary statistics. The summary statistics extracted are mean, standard deviation, minimum, maximum, and quantile values (at 0.25, 0.5, and 0.75, corresponding to the first and third quartiles and the median). The tabulated summary statistics are then fed to a random forest classifier for training a univariate time-series classifier. For applying this method to our MVTS classification task, we employed a late fusion approach, where univariate time-series classification models trained on individual parameters are ensembled based on a majority voting schema.

3.3. One-nearest Neighbor

The similarity-based (nearest neighbor) time-series classification is one of the most popular classification methods due to its simplicity, nonparametric nature, potential for interpretability, and flexibility in utilizing different similarity schemes/metrics. It assigns a label for a new instance based on the target value of k-most similar instances in the training data set. The similarity is assessed based on the inverse of a designated time-series distance measure, and utilization of elastic distance measures allows one to capture the local and global similarity patterns (J. Faouzi 2022). In this work, we implement a k-nearest neighbor (kNN, with $k = 1$) classifier with dynamic time warping (DTW) distance, which is a commonly used elastic measure (H. Sakoe & S. Chiba 1978). This 1NN is a baseline classifier for time-series classification and, similar to our earlier approaches, we use 1NN with DTW distance to issue predictions.

3.4. Supervised Time-series Forest Classifier

The STSF (N. Cabello et al. 2020) classifier comes under the category of interval-based models, and uses three representations (time, frequency, and derivative) of the input time series to extract features from a supervised selection of intervals. That is, for each representation, the STSF builds an ensemble of decision trees on intervals selected through a supervised process wherein the algorithm finds the discriminatory intervals. The STSF computes the region of interest to highlight

Table 1
Partitioning Strategy of Our Data Set

	Training	Validation	Test
Positive	80	80	84
Negative	918	894	837

Note. Number of samples in each partition corresponding to the binary target labels. Here, labels correspond to a yes (positive) or no (negative).

the location of discriminatory intervals, which are the intersected regions of such intervals. It extracts seven features, namely, the mean (μ), standard deviation (σ), slope (m), median, interquartile range, minimum value, and maximum value, from each interval. The ranking of the interval feature is obtained by a scoring function that indicates how well the feature separates one class of time series from the other classes. The final set of intervals is obtained in a top-down approach to represent the entire series. The feature set is concatenated to form a new data set upon which decision trees are built. The final output is based on majority voting of averaged probability estimates of the individual estimators in the ensemble.

3.5. Data Partitions

As we are implementing a classification task with target labels, we split our data set into three sets: the training set, the hold-out or validation set, and the test set. This partitioning schema is relevant to estimate the predictive capacity of the model on different aspects. For instance, the validation set is used in the present work to obtain the optimal threshold of classification for each model, estimate the skills scores and compare their performance. Furthermore, we use the unseen test set to sanity-check the best model's predictive capabilities.

We split our data set into three sets on nonoverlapping years (1986–1992, 1993–2002, and 2003–2018) such that there are 998 samples for training, 974 samples for validation, and 921 samples for testing the model. Here, each partition consists of a similar number of positive samples ($\approx$80) and has similar class-imbalance ratios, that is, $\approx$1:11. Table 1 shows the number of data samples in each partition in relation to the binary class labels.

4. Results

In this work, we aim to demonstrate the robustness of the methodology for SEP event forecasting that was proposed in our previous work (S. A. Rotti et al. 2024) and provide a proof of concept on its efficiency in generating short-term predictions. For this purpose, we compare three classifiers (SSC, 1NN, and STSF) and analyze their performance. We construct a binary classification task such that the "positive" class consists of strong SEP events while the "negative" class contains the rest. We use the training set to train the model and perform a grid search for hyperparameter optimization of our classifiers. The best hyperparameters for STSF and SSC were on the default model settings with the number of estimators set to 200.

We explore the model capabilities for different short-term prediction windows of 5, 15, 30, 45, and 60 minutes. We use the following terms henceforth to identify the different prediction windows that act as lead times for model predictions, respectively: T_5, T_{15}, T_{30}, T_{45}, and T_{60}. In each experiment, we consider the respective prediction window consisting of a constant 11 hr observation window from our MVTS samples before training the model. We implement multiple metrics (see Section 4.1) to estimate the predictive/forecasting skill of our models.

4.1. Metrics

We consider a 2×2 contingency table with the following elements: true positive (TP), true negative (TN), false positive (FP), and false negative (FN). Here, TP indicates the number of correctly predicted strong SEP events (positive class), while TN represents the number of correctly predicted weak and non-SEP events (negative class). FP corresponds to the number of negative classes predicted as positives (false alarms), while FN corresponds to the number of positive class labels predicted as negatives (misses).

On a simple scale, statistical metrics such as precision (Equation (1)) and recall (Equation (2)) have been traditionally used to assess classifier performance. Precision for the positive class is used to evaluate the model's correct prediction with respect to the false alarms. Recall for the positive class characterizes the ability of the classifier to find all of the positive cases:

$$\text{Precision} = \frac{(\text{TP})}{(\text{TP} + \text{FP})}, \tag{1}$$

$$\text{Recall} = \frac{(\text{TP})}{(\text{TP} + \text{FN})}. \tag{2}$$

Focusing on the importance of positive classes, we consider the F_1 score for the positive class, which can be estimated as the harmonic mean of precision and recall, as shown in Equation (3). It ranges between 0 and 1 such that scores closer to 1 indicate the model to be better:

$$F_1 = 2 \times \frac{(\text{Precision} \times \text{Recall})}{(\text{Precision} + \text{Recall})}. \tag{3}$$

To understand the overall model performance, we use a "weighted" average for the F_1 score, as shown in Equations (4) and (5), which computes the score for each target class and uses sample weights that depend on the number of instances in that class while averaging. Here, i is the number of target classes in the data set, which is two in the present work:

$$F_{1_{\text{weighted}}} = \sum_{i=1}^{N} w_i \times F_{1_i}, \tag{4}$$

$$w_i = \frac{\text{Numberofsamplesinclass}i}{\text{Totalnumberofsamples}}. \tag{5}$$

We implement true skill statistics (TSS; F. Woodcock 1976; H. Daan 1985) to account for the false-positive rate (FPR) comparing the difference between the probability of detection and the probability of false detection, as shown in Equation (6). TSS ranges from –1 to +1, where the latter indicates a perfect score. TSS $\leqslant 0$ indicates it is worse than a random classification:

$$\text{TSS} = \frac{(\text{TP} \times \text{TN}) - (\text{FP} \times \text{FN})}{(\text{TP} + \text{FN}) \times (\text{FP} + \text{TN})}. \tag{6}$$

Furthermore, we consider the Heidke skill score (HSS; P. Heidke 1926), which measures the improvement of the forecast over a random prediction, as defined in Equation (7). A HSS of 1 indicates perfect performance and 0 indicates no skill.

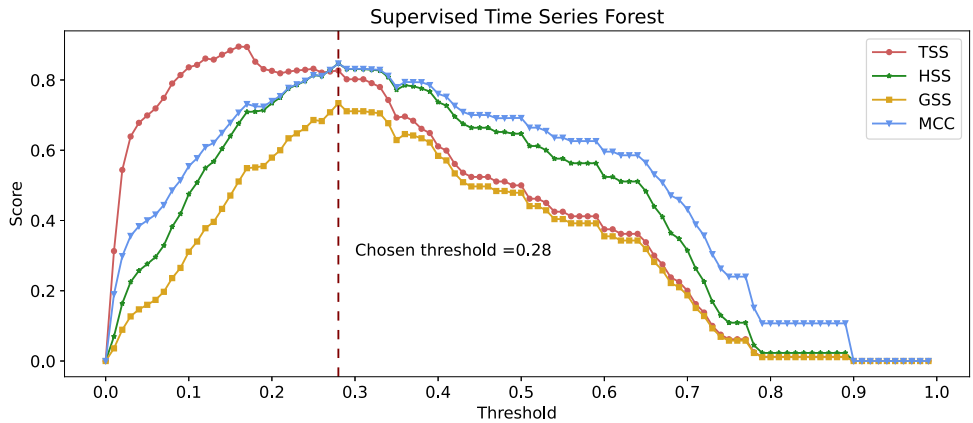

Figure 5. The variation in skills such as TSS, HSS, GSS, and MCC with respect to increasing the classification threshold for the STSF model on the validation/hold-out set at a prediction window of 30 minutes (T_{30}). The optimal threshold value for the model is indicated in the plot.

A no skill score means the forecast is not better than a random binary forecast based on class distributions:

$$\text{HSS} = \frac{2 \times ((\text{TP} \times \text{TN}) - (\text{FP} \times \text{FN}))}{((\text{TP} + \text{FN}) \times (\text{TN} + \text{FN})) + ((\text{FP} + \text{TN}) \times (\text{FP} + \text{TP}))}. \tag{7}$$

The Gilbert skill score (GSS; J. T. Schaefer 1990) considers the number of hits due to chance, which is the frequency of an event multiplied by the total number of forecast events, as given in Equation (8). The GSS ranges from –1/3 to 1. Here, 0 indicates no skill, while 1 is a perfect forecast:

$$\text{GSS} = \frac{\text{TP} - \left(\frac{(\text{TP} + \text{FN}) \times (\text{TP} + \text{FP})}{\text{TP} + \text{FP} + \text{TN} + \text{FN}}\right)}{(\text{TP} + \text{FP} + \text{FN}) - \left(\frac{(\text{TP} + \text{FN}) \times (\text{TP} + \text{FP})}{\text{TP} + \text{FP} + \text{TN} + \text{FN}}\right)}. \tag{8}$$

Accounting for TNs to assess the performance of a binary class problem is essential in our context. Hence, we also implement Matthew's correlation coefficient (MCC), as defined in Equation (9). MCC ranges from –1 to 1, where 0 indicates no skill and 1 shows perfect agreement between predicted and actual values. MCC is a robust metric that follows a comprehensive and balanced strategy for using contingency table elements. That is, a high MCC score is obtained only when the predictions proportionately account for the size of both positive and negative samples in the data set. This is important because popular scores such as the F_1 score fall short in accounting for class imbalance:

$$\text{MCC} = \frac{(\text{TP} \times \text{TN}) - (\text{FP} \times \text{FN})}{\sqrt{(\text{TP} + \text{FP}) \times (\text{TP} + \text{FN}) \times (\text{TP} + \text{FP}) \times (\text{TN} + \text{FN})}}. \tag{9}$$

4.2. Validation Set

As mentioned in Section 3.5, we use the second partition from our data set as a hold-out portion to estimate the optimal threshold for our classifiers. The classification threshold is the decision threshold that allows us to map the probabilistic output of a binary classifier to a binary category. In other words, it is a cutoff point used to assign a specific predicted class label for each sample. By default, the classification threshold in our models is 0.5. What that means is any prediction above 0.5 belongs to the positive class and that below 0.5 belongs to the negative class. However, 0.5 is not always optimal, especially when we have an imbalanced data set. Therefore, we identify a reliable threshold for the classifier that better discriminates between the two class labels. The first step is to derive the prediction probabilities from the validation set and extract the best trade-off between true-positive rate (TPR) and FPR. In general, any binary classification model predicts mean probabilities for each input sample belonging to the positive class, where the prediction score from the classifier is greater than a parameterized threshold. Then, a classification threshold (from 0.0 to 1.0) is used to assign a binary label to the predicted probabilities:

$$J = \text{Sensitivity} + \text{Specificity} - 1. \tag{10}$$

To find the optimal threshold that minimizes the difference between TPR and FPR, the Youden index (J; W. J. Youden 1950) is used in the literature, as defined in Equation (10). Here, sensitivity is the recall for the positive class and specificity is the recall for the negative class. J is the specialized version of TSS for binary classification tasks. As a characteristic example, we show in Figure 5 the effect of "thresholding" on the performance of STSF by visualizing the variations in the skills due to changing thresholds. Here, the estimated optimal threshold of STSF for T_{30} on our data generation and sampling method is 0.28. We obtain optimal classification thresholds at all prediction windows for all the models and estimate the model skills. A comparison of the models' performance is shown in Figure 6.

Computationally, the runtime of SSC was the fastest, followed by STSF and then 1NN. In terms of performance, 1NN falls behind both models but maintains relatively close values for all the skill scores in each prediction window with $\sigma \approx 0.04$. On the other hand, SSC shows a similar trend in its skills with $\sigma \approx 0.01$ except for TSS ($\sigma \approx 0.07$). Although SSC was a little faster, the performance marginally lags on our data set compared to STSF. For example, the average TSS for the SSC is ≈14% lower than that for the STSF. The STSF model performs exceptionally well compared to the other two baseline classifiers at all parameters. Specifically, the high scores at T_{60} (e.g., TSS = 0.850) provide satisfactory reliability to our methodology. Table 2 presents contingency tables for all prediction windows corresponding to the optimal threshold of STSF on the validation set. The contingency tables for SSC and 1NN are provided in Tables 4 and 5, respectively, in the Appendix.

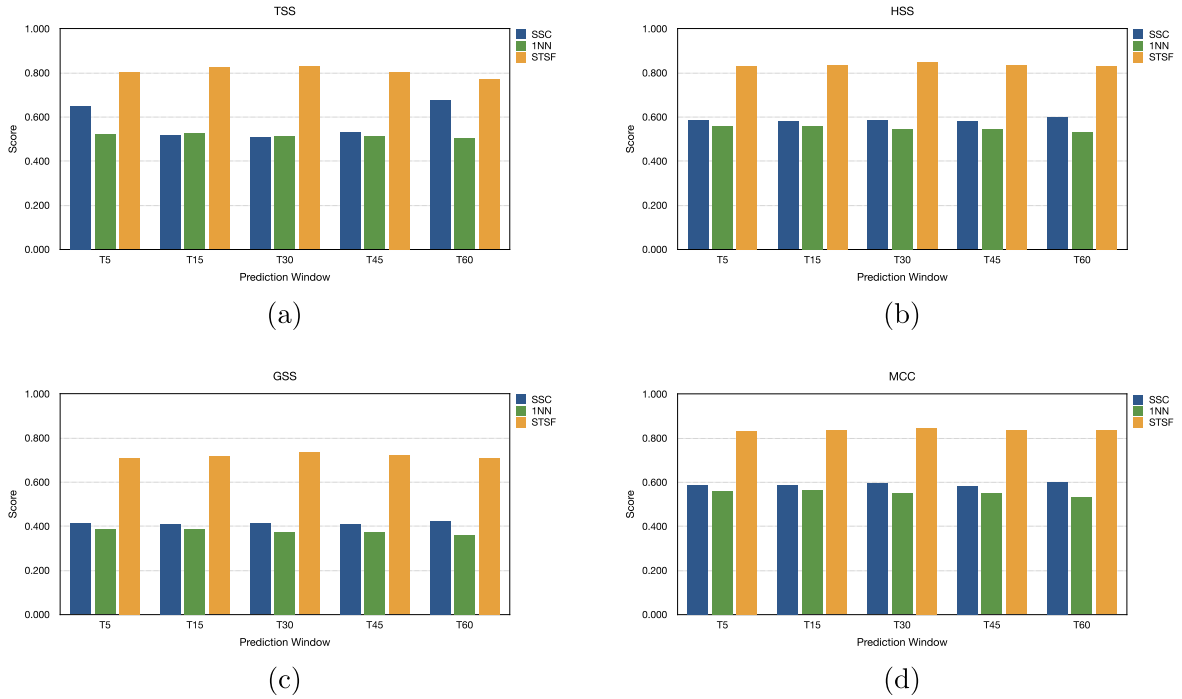

Figure 6. Comparison of the SSC, STSF, and 1NN classifiers with respect to variation in skills such as (a) TSS, (b) HSS, (c) GSS, and (d) MCC for prediction windows T_5, T_{15}, T_{30}, T_{45}, and T_{60} on the validation set.

Table 2
Contingency Tables for the STSF Model on the Validation Set

		T_5		T_{15}		T_{30}		T_{45}		T_{60}	
		Predicted		Predicted		Predicted		Predicted		Predicted	
		Strong	Weak	Strong	Weak	Strong	Weak	Strong	Weak	Strong	Weak
True	Strong	65	15	67	13	67	13	65	15	62	18
	Weak	9	885	11	883	9	885	8	886	5	889

Notes. Truth tables for the STSF model for different prediction windows. The first column is a shared entry of true labels against predictive labels for each experiment. The elements indicate the number of predictions with respect to the actual occurrences in the validation set. Prediction windows are shown by a subscript to T. For example, T_5 indicates a prediction window of 5 minutes. Model name: supervised time-series forest (STSF).

4.3. Test Set

In our previous work (S. A. Rotti et al. 2024), we found that an ensemble of interval-based STSF classifiers is a highly efficient and optimal model to predict strong SEP events. STSF's approach of computing interval features in a supervised manner from three different time-series representations by using robust statistics was a viable option for our problem specification. In this paper, we extend our modeling strategy by including non-SEP periods in our data set. Furthermore, we compare the performance of STSF and two baseline classifiers, namely, SSC and 1NN. Based on the forecasting skill scores, we use the validation set from our data partitions to estimate the models' performance. This shows that STSF is the best model under all comparative parameters. We use the test set to further assess the model's capabilities.

In Table 3, we show the contingency table, or the confusion matrix, elements derived for the STSF classifier on the test set for all prediction windows (T_5 to T_{60}) considered in this work. The increasing prediction window shows a relatively consistent type II error (false-negative rate) except for T_{30}. That is, the number of false positives is lower compared to false negatives. A possible reason for the situation here is that several weak SEP events have proton fluxes closely fluctuating in the vicinity of the SWPC "S1" threshold. Detecting such patterns at high precision becomes relevant to reducing FNs.

In Figure 7, we compare the skill scores of STSF with respect to the different prediction windows considered in this work. The average TSS at all prediction windows is $\approx$0.86 ($\pm$0.02). Similarly, the HSS, MCC, and GSS scores obtained are also high. There is a slight increment in the scores from T_5 to T_{30}, but this is reduced at T_{45} and T_{60}. Such slight variation is due to subtle changes in the optimal classification threshold $\approx$0.3($\pm$0.03) for the STSF. The scores for T_{60} only have a marginal reduction ($<$2%) compared to T_5. Furthermore, there is a decrease of $\approx$7%($\pm$2%) in skill scores compared to S. A. Rotti et al. (2024). This is expected due to the extended lead times considered with a larger class imbalance in the data set. Overall, the classification and computational efficiencies we obtain for STSF are highly satisfactory and viable. This means that our work adequately provides a proof of concept of our modeling architecture for short-term predictive capabilities up to a 60 minutes window and can be further transformed into implementation for NRT operations.

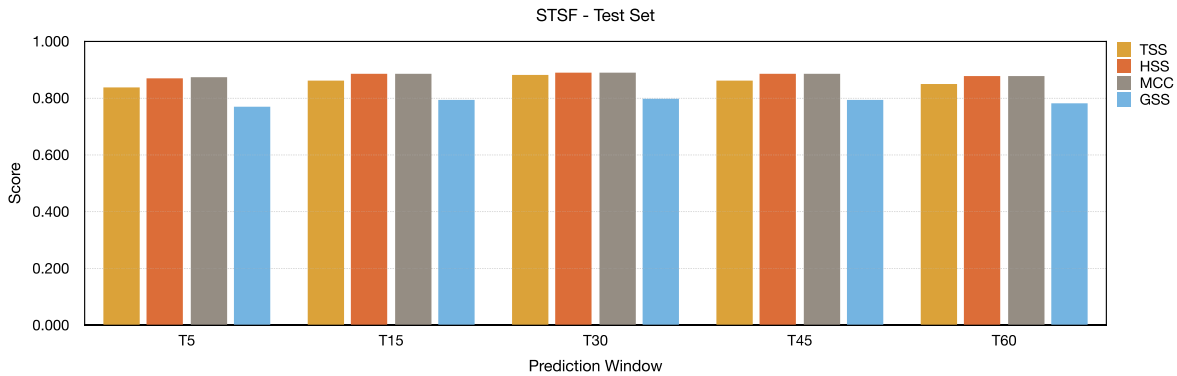

Figure 7. Comparison of skill scores of the STSF model on the test set at prediction windows T_5, T_{15}, T_{30}, T_{45}, and T_{60}.

Table 3
Contingency Tables for the STSF Model on the Test Set

		T_5		T_{15}		T_{30}		T_{45}		T_{60}	
		Predicted		Predicted		Predicted		Predicted		Predicted	
		Strong	Weak	Strong	Weak	Strong	Weak	Strong	Weak	Strong	Weak
True	Strong	71	13	73	11	75	9	73	11	72	12
	Weak	6	831	6	831	8	829	6	831	6	831

Notes. Truth tables for the STSF model for different prediction windows. The first column is a shared entry of true labels against predictive labels for each experiment. The elements indicate the number of predictions with respect to the actual occurrences in the test set. Prediction windows are shown by a subscript to T. For example, T_5 indicates a prediction window of 5 minutes. Model name: supervised time-series forest (STSF).

5. Conclusions

In this work, we approach the SEP event short-term prediction problem from a time-series classification perspective. We primarily use the GSEP data set but extend our negative sample size by including "SEP-quiet" periods. That is, we consider ⩾C6.0 flares that do not lead to SEP events observed near Earth. Our data set constitutes in situ time-series measurements from the NOAA GOES missions for solar cycles 22 to 24. We utilize the long-band (1–8 Å) X-ray measurements from the XRS instrument and proton fluxes (P3, P5, and P7) from the SEM instrument on board the GOES missions. Our data set has a total of 2893 samples, of which 244 are strong SEP events. We defined the target labels based on the SWPC "S1" threshold for an SEP event. Positive labels are strong events crossing the 10 pfu threshold in the 10 MeV proton channel, and negative otherwise.

Our goal is to develop robust models that can successfully find discriminants between SEP and SEP-quiet patterns. In our recent study (S. A. Rotti et al. 2024), we found our methodology of using an ensemble of feature-based univariate time-series classifiers to perform very well in classifying between strong and weak SEP events. Specifically, we found that the STSF (N. Cabello et al. 2020) classifier obtained optimal results. In the present work, we implement our earlier methodology on an extended data set and compare the performance of three models, namely, the SSC, 1NN, and STSF classifiers. Furthermore, we consider a fixed observational window of 11 hr and use multiple prediction windows with a prediction interval of 5, 15, 30, 45, and 60 minutes to leverage on high confidence for relatively low lead times.

In our model analysis phase, we used the Youden index (J) to understand the trade-off between the TPR and the FPR at different classification thresholds. That is, we estimate an optimal threshold that provides a TPR with an acceptable FPR to make decisions using our classifiers. We utilized the validation set to compare the performance of the models based on multiple skill scores, such as the TSS, HSS, GSS, and MCC. This resulted in STSF performing better compared to SSC and 1NN at all prediction windows. Furthermore, we use an unseen test set to obtain a classification report of the best model, including contingency tables. For a 60 minutes prediction window, we obtain the following scores for STSF: TSS = 0.850, HSS = 0.878, GSS = 0.783, and MCC = 0.879. In summary, our results are promising and provide a gateway for our model architecture to be used in an operational context to offer short-term predictions on the arrival of strong SEP events up to 60 minutes. The data set and coding methodology of our model implementation have been made publicly available on the Harvard Dataverse (doi:10.7910/DVN/MWY6H7) under a Creative Commons license.

The applications in which we foresee our predictive system being useful to space agencies across the globe are twofold. First, the system can serve as a convenient tool for forecasters, for example at NASA's Space Radiation Analysis Group and NOAA's SWPC, in assisting them in broadcasting highly accurate last-minute warnings to astronauts on the surface of the Moon or during extravehicular activities when outside the Earth's magnetosphere. Second, for future interplanetary missions, such as the missions to Mars that NASA is planning, the spacecraft will spend most of its time outside of the Sun-Earth line. Therefore, SEP predictions that apply to the Earth-Moon system will be of limited value when the spacecraft-Sun makes a large angle with the Earth-Sun direction. A crewed spacecraft on a journey to Mars would only have to carry a scaled-down version of the GOES instruments to produce short-term warnings of intense solar particle storms. Hence, we

expect that our system can be a valuable addition to the tools that space agencies will need to protect astronauts on upcoming lunar and Mars missions.

Acknowledgments

We acknowledge the use of X-ray and proton flux data from the GOES missions made available by NOAA. We thank the anonymous reviewer for constructive comments on the manuscript that have improved the contents of the paper. Authors P. M.'s and B.A.'s contribution to this work is supported by NASA SWR2O2R grant No. 80NSSC22K0272. Author S.R. carried out this work while supported by NASA FINESST grant No. 80NSSC21K1388 and SMD grant No. 24-SMDSS24-0045.

Software: pandas (W. McKinney et al. 2010), numpy (S. Van Der Walt et al. 2011; C. R. Harris et al. 2020), sklearn (F. Pedregosa et al. 2011), sktime (M. Löning et al. 2019; M. Löning et al. 2022), matplotlib (J. D. Hunter 2007), cartopy (Met Office 2010-2015).

Appendix
Contingency Tables

The contingency table elements for SSC and 1NN on the validation set are provided in Tables 4 and 5, respectively, for each of our experiments on changing prediction windows. The values from these tables are used to estimate the skill scores that are defined in Section 4.1 and shown in Figure 6.

Table 4
Contingency Tables for the SSC Model on the Validation Set

		T_5		T_{15}		T_{30}		T_{45}		T_{60}	
		Predicted		Predicted		Predicted		Predicted		Predicted	
		Strong	Weak	Strong	Weak	Strong	Weak	Strong	Weak	Strong	Weak
True	Strong	56	24	43	18	42	38	44	36	58	22
	Weak	44	850	37	876	15	879	20	874	45	849

Notes. Truth tables for the SSC model for different prediction windows. The first column is a shared entry of true labels against predictive labels for each experiment. The elements indicate the number of predictions with respect to the actual occurrences in the validation set. Prediction windows are shown by a subscript to T. For example, T_5 indicates a prediction window of 5 minutes. Model name: summary statistic classifier (SSC).

Table 5
Contingency Tables for the 1NN Model on the Validation Set

		T_5		T_{15}		T_{30}		T_{45}		T_{60}	
		Predicted		Predicted		Predicted		Predicted		Predicted	
		Strong	Weak	Strong	Weak	Strong	Weak	Strong	Weak	Strong	Weak
True	Strong	44	36	44	36	43	37	43	37	43	37
	Weak	25	869	24	870	25	869	25	869	29	865

Notes. Truth tables for the 1NN model for different prediction windows. The first column is a shared entry of true labels against predictive labels for each experiment. The elements indicate the number of predictions with respect to the actual occurrences in the validation set. Prediction windows are shown by a subscript to T. For example, T_5 indicates a prediction window of 5 minutes. Model name: one-nearest neighbor (1NN).

ORCID iDs

Sumanth A. Rotti https://orcid.org/0000-0003-1080-3424
Berkay Aydin https://orcid.org/0000-0002-9799-9265
Petrus C. Martens https://orcid.org/0000-0001-8078-6856

References

Alberti, T., Laurenza, M., Cliver, E. W., et al. 2017, ApJ, 838, 59
Aminalragia-Giamini, S., Raptis, S., Anastasiadis, A., et al. 2021, JSWSC, 11, 59
Anastasiadis, A., Papaioannou, A., Sandberg, I., et al. 2017, SoPh, 292, 134
Angryk, R. A., Martens, P. C., Aydin, B., et al. 2020, NatSD, 7, 227
Arbib, M. A. 2003, The Handbook of Brain Theory and Neural Networks (Cambridge, MA: MIT Press)
Bain, H., Steenburgh, R., Onsager, T., & Stitely, E. 2021, SpWea, 19, e2020SW002670
Balch, C. C. 2008, SpWea, 6, S01001
Beck, P., Latocha, M., Rollet, S., & Stehno, G. 2005, AdSpR, 36, 1627
Bornmann, P. L., Speich, D., Hirman, J., et al. 1996, Proc. SPIE, 2812, 291
Boubrahimi, S. F., Aydin, B., Martens, P., & Angryk, R. 2017, in 2017 IEEE Int. Conf. on Big Data (Big Data) (Piscataway, NJ: IEEE), 2533
Cabello, N., Naghizade, E., Qi, J., & Kulik, L. 2020, in 2020 IEEE Int. Conf. on Data Mining (ICDM) (Piscataway, NJ: IEEE), 948
Camporeale, E. 2019, SpWea, 17, 1166
Cane, H. V., McGuire, R. E., & von Rosenvinge, T. T. 1986, ApJ, 301, 448
Cane, H. V., Reames, D. V., & von Rosenvinge, T. T. 1988, JGR, 93, 9555
Chen, P. 2011, LRSP, 8, 1
Cliver, E. W., & D'Huys, E. 2018, ApJ, 864, 48
Cliver, E. W., Schrijver, C. J., Shibata, K., & Usoskin, I. G. 2022, LRSP, 19, 2
Daan, H. 1985, in Probability, Statistics, And Decision Making In The Atmospheric Sciences, ed. A. H. Murphy & R. W. Katz (Boca Raton, FL: CRC Press), 379
Desai, M., & Giacalone, J. 2016, LRSP, 13, 3
Dierckxsens, M., Tziotziou, K., Dalla, S., et al. 2015, SoPh, 290, 841
Engell, A. J., Falconer, D. A., Schuh, M., Loomis, J., & Bissett, D. 2017, SpWea, 15, 1321
Falconer, D., Barghouty, A. F., Khazanov, I., & Moore, R. 2011, SpWea, 9, S04003
Faouzi, J. 2022, Machine Learning, Emerging Trends and Applications (Vilnius: Diamond Scientific Publishing)
Feynman, J., & Hundhausen, A. J. 1994, JGR, 99, 8451
Garcia, H. A. 1994, SoPh, 154, 275
Gopalswamy, N., Lara, A., Yashiro, S., Kaiser, M. L., & Howard, R. A. 2001, JGR, 106, 29207
Gopalswamy, N., Mäkelä, P., Yashiro, S., et al. 2017, JPhCS, 900, 012009
Gopalswamy, N., Yashiro, S., Xie, H., et al. 2008, ApJ, 674, 560
Grubb, R. N. 1975, The SMS/GOES Space Environment Monitor Subsystem, Tech. Memorandum ERL SEL-42, NOAA, https://repository.library.noaa.gov/view/noaa/18586
Hansen, L. K., & Salamon, P. 1990, ITPAM, 12, 993
Harris, C. R., Millman, K. J., Van Der Walt, S. J., et al. 2020, Natur, 585, 357
Heidke, P. 1926, Geografiska Annaler, 8, 301
Hosseinzadeh, P., Filali Boubrahimi, S., & Hamdi, S. M. 2024, ApJS, 270, 31
Hunter, J. D. 2007, CSE, 9, 90
Jackman, C. H., & McPeters, R. D. 1987, PhST, 1987, 309
Jiggens, P., Clavie, C., Evans, H., et al. 2019, SpWea, 17, 99
Kahler, S. W. 1992, ARA&A, 30, 113
Kahler, S. W., Cliver, E. W., & Ling, A. G. 2007, JASTP, 69, 43
Klein, K.-L., & Dalla, S. 2017, SSRv, 212, 1107
Laurenza, M., Cliver, E. W., Hewitt, J., et al. 2009, SpWea, 7, S04008
Lavasa, E., Giannopoulos, G., Papaioannou, A., et al. 2021, SoPh, 296, 107
Löning, M., Bagnall, A., Ganesh, S., et al. 2019, arXiv:1909.07872
Löning, M., Király, F., Bagnall, T., et al. 2022, sktime/sktime: v0.13.4, Zenodo, doi:10.5281/zenodo.7117735
Low, B. C. 1996, SoPh, 167, 217
Marqué, C., Posner, A., & Klein, K.-L. 2006, ApJ, 642, 1222
McKinney, W. 2010, Data Structures for Statistical Computing in Python, 445, 51
Met Office 2010-2015, Cartopy: A Cartographic Python Library with a Matplotlib Interface, Exeter, Devon, https://scitools.org.uk/cartopy
Milligan, R. O., & Ireland, J. 2018, SoPh, 293, 18
Nunez, M. 2011, SpWea, 9, S07003
Onsager, T., Grubb, R., Kunches, J., et al. 1996, Proc. SPIE, 2812, 281
Papaioannou, A., Anastasiadis, A., Kouloumvakos, A., et al. 2018, SoPh, 293, 100
Parker, E. N. 1965, SSRv, 4, 666
Pedregosa, F., Varoquaux, G., Gramfort, A., et al. 2011, JMLR, 12, 2825
Posner, A. 2007, SpWea, 5, 05001
Reames, D. V. 1999, SSRv, 90, 413
Reames, D. V. 2013, SSRv, 175, 53
Rotti, S., Aydin, B., Georgoulis, M., & Martens, P. 2022a, GSEP Dataset, V5, Harvard Dataverse, doi:10.7910/DVN/DZYLHK
Rotti, S., Aydin, B., Georgoulis, M. K., & Martens, P. C. 2022b, ApJS, 262, 29
Rotti, S., & Martens, P. C. 2023, ApJS, 267, 40
Rotti, S. A., Aydin, B., & Martens, P. C. 2024, ApJ, 966, 165
Rotti, S. A., Martens, P. C. H., & Aydin, B. 2020, ApJS, 249, 20
Sakoe, H., & Chiba, S. 1978, ITASS, 26, 43
Sauer, H. H. 1989, in AIP Conf. Proc. 186, High-Energy Radiation Background in Space (Melville, NY: AIP), 216
Schaefer, J. T. 1990, WtFor, 5, 570
Schapire, R. E. 1990, Machine Learning, 5, 197
Schrijver, C. J., & Siscoe, G. L. 2010, Heliophysics: Space Storms and Radiation: Causes and Effects (Cambridge: Cambridge Univ. Press)
Schwadron, N. A., Townsend, L., Kozarev, K., et al. 2010, SpWea, 8, 00E02
Smart, D. F., & Shea, M. A. 1992, AdSpR, 12, 303
Swalwell, B., Dalla, S., & Walsh, R. W. 2017, SoPh, 292, 173
Usoskin, I. G. 2017, LRSP, 14, 3
Van Der Walt, S., Colbert, S. C., & Varoquaux, G. 2011, CSE, 13, 22
Van Hollebeke, M. A. I., Ma Sung, L. S., & McDonald, F. B. 1975, SoPh, 41, 189
Whitman, K., Egeland, R., Richardson, I. G., et al. 2022, AdSpR, 72, 5161
Woodcock, F. 1976, MWRv, 104, 1209
Youden, W. J. 1950, Cancer, 3, 32