
Article:
Francisco, G. orcid.org/0000-0003-3694-7813, Berretti, M. orcid.org/0009-0007-2465-1931, Chierichini, S. orcid.org/0009-0005-6746-2917 et al. (4 more authors) (2025) Limits of solar flare forecasting models and new deep learning approach*. The Astrophysical Journal, 985 (1). 108. ISSN 0004-637X

https://doi.org/10.3847/1538-4357/adc56d

The Astrophysical Journal, 985:108 (17pp), 2025 May 20 https://doi.org/10.3847/1538-4357/adc56d

Limits of Solar Flare Forecasting Models and New Deep Learning Approach*

G. Francisco[1,2,3], M. Berretti[1,4], S. Chierichini[1,3,5], R. Mugatwala[1,3,5], J. Fernandes[6], T. Barata[2], and D. Del Moro[1]
[1] Department of Physics, University of Rome Tor Vergata, Rome, Italy
[2] Instituto de Astrofisica e Ciências do Espaço (IA), University of Coimbra, Coimbra, Portugal
[3] Department of Physics, University of Rome La Sapienza, Rome, Italy
[4] University of Trento, Trento, Italy
[5] SP2RC, School of Mathematics and Statistics, University of Sheffield, Sheffield, UK
[6] CITEUC, Geophysical and Astronomical Observatory, University of Coimbra, Department of Mathematics, Coimbra, Portugal
Received 2024 September 3; revised 2025 March 21; accepted 2025 March 23; published 2025 May 16

Abstract

Reliable forecasting models are necessary to mitigate the risks posed by solar flares to human technology. This study introduces a novel deep learning forecasting approach while emphasizing the need for performance evaluation methods tailored to better highlight current models' limitations. In particular, we show that models reaching state-of-the-art performance with traditional metrics have similar explanatory power to no-skill persistence models and notably struggle to forecast change in activity significantly better than random guesses. We also discuss shortcomings in traditional evaluation metrics like the True Skill Statistic (TSS), which we show to be mathematically dependent on the class balance for specific models. We introduce patch-distributed CNNs, which allow us to perform full-disk forecasts while providing event probabilities in solar subregions and position predictions. This new framework offers similar information to active region (AR)-based forecasting models while bypassing the problem of unrecorded and misattributed flares that are detrimental to machine learning training. As a result, the model also operates independently of prior feature extraction and AR detection, thus offering promising operational utility with minimal external dependencies. Finally, a method is proposed for constructing balanced and independent cross-validation folds for full-disk models. Models combining Solar Dynamic Observatory (SDO)/Atmospheric Imaging Assembly EUV images as inputs show improved performance compared to employing SDO/HMI photospheric magnetograms, with a TSS of 0.74 for the C+ model and 0.62 for the M+ model.

Unified Astronomy Thesaurus concepts: Solar flares (1496); Convolutional neural networks (1938); Magnetogram (2359)

1. Introduction

1.1. Background and Related Works

Solar flares are one of the most energetic manifestations of solar activity. They are bursts of electromagnetic radiation and particles believed to be caused by magnetic reconnections converting huge amounts of magnetic energy into heat and kinetic energy. To characterize their potential danger, flares are commonly classified according to their Soft X-Rays (SXRs; wavelength from 0.1 to 0.8 nm) maximum peak flux (MPF): A-class flares with a MPF $< 10^{-7}\,\mathrm{W\,m^{-2}}$, B-class flares with MPF $\in [10^{-7}, 10^{-6}[\mathrm{W\,m^{-2}}$, C-class flares with MPF $\in [10^{-6}, 10^{-5}[\mathrm{W\,m^{-2}}$, M-class flares with MPF $\in [10^{-5}, 10^{-4}[\mathrm{W\,m^{-2}}$, and X-class flares with MPF $> 10^{-4}\,\mathrm{W\,m^{-2}}$. Flares above the M-class start representing a threat to human health and technologies, motivating efforts toward reliable forecasting methods.

Early machine learning solutions typically began by engineering physically interpretable features—often derived from magnetogram observations—believed to correlate with imminent flare activity. These features were then used as inputs to train predictive models. However, more recent approaches exploit deep learning methods that automatically learn features from images or time series data. For example, X. Huang et al. (2018), E. Park et al. (2018), X. Li et al. (2020), Z. Deng et al. (2021), and C. Pandey et al. (2023) employed convolutional neural network (CNN) architectures on magnetogram images, while N. Nishizuka et al. (2018) used a multi-layer perceptron (MLP) artificial neural network on a combination of physical parameters. Other works combined CNN and long short-term memory (LSTM) modules to capture both spatial and temporal information (S. Guastavino et al. 2022a; Z. Sun et al. 2022), or combined CNN with traditional techniques such as random forests (V. Deshmukh et al. 2022).

1.1.1. Current Limitations

Although these works have advanced the field, major limitations persist. First, forecasting flare activity directly on preidentified active regions (ARs) leads to a dependence on external AR-detection procedures, which can miss or misattribute events. K. Van der Sande et al. (2022) noted up to 8% misattributed labels and about 20% missing M-class or above flares in standard catalogs, highlighting potential errors that degrade both training and evaluation. Second, most full-disk-level models (E. Park et al. 2018; K. Yi et al. 2021; C. Pandey et al. 2023) do not provide information about flare positions, limiting their operational usefulness when event localization is required. Third, model evaluations in existing studies often rely on standard machine learning metrics and conventional train-

* V1 preprint released on 2024 February, V2 released on 2024 May 15 on ESS Open Archive 10.22541/essoar.170688972.24631782/v3.

test splits, which, while widely used, may be inadequate for this specific task due to data set selection biases and the dynamic nature of flare production (G. Barnes et al. 2016; K. D. Leka et al. 2019; T. Cinto et al. 2020; S. Guastavino et al. 2022a, 2022b).

1.1.2. Present Work's Contributions

This paper aims to address these limitations by introducing:

1. A weakly supervised learning framework with a full-disk forecasting model using a CNN that does not rely on AR identification. Unlike previous full-disk approaches, it retrieves subregional predictions and potential flare positions while using only full-disk labels at training, thus mitigating issues arising from misallocated flares or unrecorded ARs.
2. A robust evaluation procedure that includes new metrics to quantify improvements over a simple persistence baseline and to gauge performance on periods exhibiting changes of solar activity. These methods reveal that, while models can appear powerful under conventional metrics, they may actually struggle to surpass basic no-skill benchmarks when predicting activity changes.
3. A careful cross-validation (CV) strategy that reduces bias in training and testing in full-disk context, leading to more realistic performance estimates suitable for real operational contexts.

These contributions together aim to advance solar flare forecasting by minimizing labeling errors, improving autonomy from external modules, and demonstrating more comprehensive evaluation procedures that help reflect operational utility.

We employ the flare catalog from N. Plutino et al. (2023), focusing on 24 hr forecasting windows. Throughout this work, a model predicting whether at least one flare above the C-class threshold will occur in the next 24 hr is referred to as a C+ model, while predicting at least one flare above the M-class threshold is referred to as an M+ model.

This work is organized as follows. Section 2 reviews commonly used binary metrics, highlighting their limitations (Section 2.1), and then introduces complementary evaluation methods (Section 2.2). Section 3 introduces our methodology, presenting the data (Section 3.1), our weakly supervised model (Section 3.2), and our training and new evaluation procedures (Section 3.3). Section 4 presents the main results, and Section 5 discusses the remaining limitations and orientations for future work.

2. Metrics

This section provides an overview of commonly employed metrics in flare forecasting, some of their drawbacks, and new evaluation methods that better highlight models' limitations—particularly concerning changes in solar activity. For conciseness and clarity, we place basic formulas and additional discussions in Appendix A.

2.1. Standard Binary Evaluation Methods

Flare forecasting typically deals with highly imbalanced data, posing challenges for model evaluation. Common metrics are defined through the confusion matrix (Equation (A1)), whose elements true positive (TP), true negative (TN), false positive (FP) and false negative (FN) can be translated into class accuracy rates (e.g., true positive rate (TPR), true negative rate (TNR)) and class precision rates (e.g., positive predictive value (PPV), negative predictive value (NPV)). In practical "alert system" scenarios, the false alarm ratio (FAR; the complement of PPV) is also closely monitored to mitigate false alarms.

F1 score. Defined in Equation (A6), the F1 score is the harmonic mean of PPV and TPR. It provides a single-value assessment for alert systems, balancing detection of positive events against excessive false alarms. Variants like $F\beta$-score can assign different weights to these two quantities, reflecting operational priorities.

True skill statistic (*TSS*). The TSS (Equation (A7)) is widely used in flare forecasting and is independent of the class ratio only under very specific assumptions (Appendix A.2.1). In practice, many flare forecasting models exhibit strong sensitivities to the class balance and the rate of changes in activity, causing the TSS to vary with data set composition. Moreover, the TSS includes no information about PPV, making it incomplete and potentially misleading in imbalanced situations (Appendix A.2.2).

Heidke Skill Score (*HSS*). The HSS (Equation (A10)) compares model performance to random guessing and accounts for both class accuracy and class precision rates. However, it may be difficult to interpret or compare across models because it encompasses TSS and markedness (Equation (A8)) with weights that depend on the model's frequency bias (Appendix A.1.7).

Matthews correlation coefficient (*MCC*). The MCC (Equation (A11)) is a less common, but often more robust, measure of a model's explanatory power. It treats positives and negatives symmetrically, synthesizing all four confusion matrix components (TP, TN, FP, FN) in a balanced way (Appendix A.2.4). It also tends to be more stable with respect to variations in data set composition (Appendix B).

Ultimately, no single metric can fully capture all the strengths and weaknesses of a model. For alert systems, the F1 score (or $F\beta$) is particularly relevant, while the MCC provides a reliable and agnostic measure of a model's explanatory power in imbalanced cases.

Thresholding. Many flare forecasting studies optimize their decision threshold to maximize a selected metric. However, K. D. Leka et al. (2019) showed that threshold "optimality" often depends greatly on the class balance of the data set, introducing additional biases in performance estimates. To avoid this complication, we evaluate all metrics at a fixed threshold of 0.5.

2.2. Identifying Flare Forecasting Models' Weaknesses

Although standard metrics offer valuable insights, they can mask key deficiencies when forecasting changes in solar activity. We therefore introduce complementary evaluations to better pinpoint where models fail.

2.2.1. Activity-change and No-change Performances

We label each time window as activity change (AC) if its binary flare label differs from the previous consecutive, non-overlapping window; otherwise, we classify it as no change (NC). Figure 1 illustrates examples of such windows in the case of M+ forecasting. Models can achieve decent performance on NC windows—basically "recognizing" a stable configuration

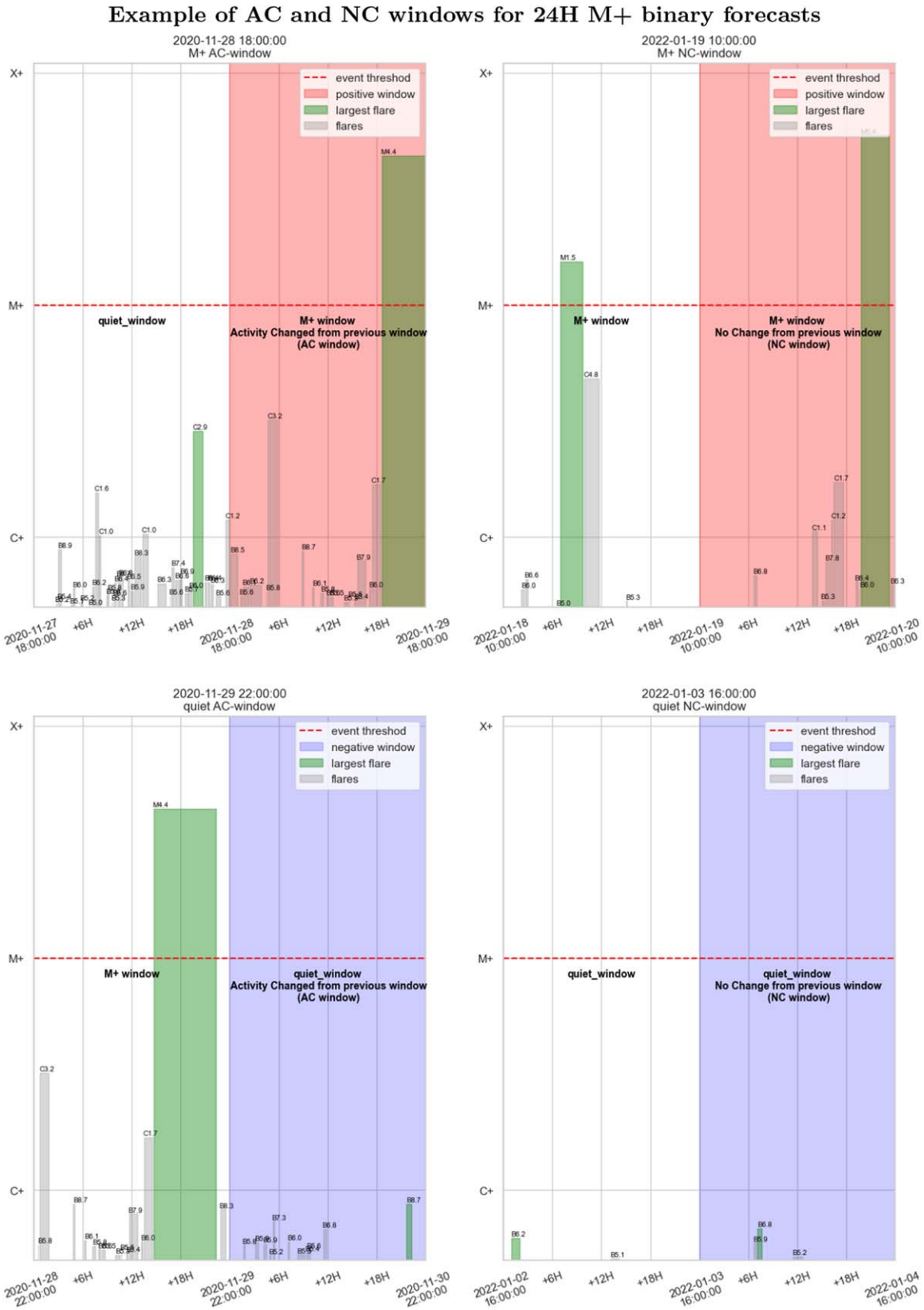

Figure 1. The first column displays examples of AC windows, with a positive AC window on the first row (colored in red) and a negative AC window on the second row (colored in blue). The second column displays examples of NC windows: positive NC window on the first row and negative NC window on the second one. The white time windows are the ones that precede the windows of interest. The gray bars correspond to flares plotted from their starting to end dates. The green bars correspond to the biggest flare inside the corresponding window. The label on top of the flares corresponds to their SXR-MPF.

of the solar activity—yet struggle on AC windows, which require the crucial ability to forecast the first flare after a quiet period and the first quiet period after an active one. This limitation is often not visible from the performance evaluated on the whole data set with standard metrics like HSS and TSS, which can reach high scores. To highlight these shortcomings, we recommend evaluating metrics separately on AC and NC subsets, denoted AC metrics and NC metrics, respectively.

2.2.2. Persistence Relative Skill Scores

Common skill scores compare model performances to random or constant baselines. In flare forecasting, a persistence model predicting that the next time window's label will be the same as the current one and can be surprisingly competitive (Section 4.1 and Appendix B.2). We define the persistence relative skill score (PRSS) (Equation (1)) that rescales a metric relative to the persistence model, producing a value in $[-1, 1]$ to quantify the degree to which a model over- or underperforms the no-skill persistence.

$$\mathrm{PRSS} = \begin{cases} \dfrac{S_{\mathrm{model}} - S^*}{S_{\mathrm{up}} - S^*} & \text{if } S_{\mathrm{model}} \geqslant S^* \\ \dfrac{S_{\mathrm{model}} - S^*}{S^* - \mathrm{In}\, f} & \text{if } S_{\mathrm{model}} < S^* \end{cases} \tag{1}$$

This normalization often reduces sensitivity to data set composition biases and gives a clearer picture of real-world utility, particularly for imbalanced tasks (Appendix B.2), where S_{model} represents the model's score, S^* is the score of the persistence baseline, and $\mathrm{S_{up}}$ and $\mathrm{In}f$ denote the upper and lower theoretical limits of the metric being normalized. By combining standard metrics such as F1 and MCC with AC/NC breakdowns and persistence relative scores, we establish a more comprehensive evaluation framework. This approach helps identify practical shortcomings, particularly in forecasting transitions between active and quiet periods

3. Method

3.1. Data

We prepared a data set of 56,664 samples spanning 2010 May 14–2023 April 18, at a temporal cadence of 2 hr. The interval from 2010 May to 2019 December serves as the training and CV period, while the interval from 2020 January to 2023 April constitutes the test data set. This temporal split ensures that the test data, drawn from a distinct solar cycle, meets the criteria of an "operational" test set as defined per T. Cinto et al. (2020), providing a reliable evaluation of generalization performance at the start of a new solar cycle.

Each sample comprises:

1. One SDO/HMI (W. D. Pesnell et al. 2012) photospheric line-of-sight (LOS) magnetogram, and
2. A triple-channel image combining the 193 Å, 211 Å, and 94 Å SDO/AIA (J. R. Lemen et al. 2012) extreme ultraviolet (EUV) observations of the solar corona.

We selected these three coronal wavelengths as they collectively cover a broad range of plasma temperatures in the solar corona. Specifically, the 211 Å channel is highly sensitive to medium-intensity activity around 2MK (typical of AR), the 94 Å channel is more responsive to hotter plasma at approximately 6MK (often associated with flares), and the 193 Å channel can observe both cooler plasma near 1MK and extremely hot plasma reaching 20MK (P. S. Athiray & A. R. Winebarger 2024). This choice provides rich thermal information of the corona and is thus suitable for our aim to compare magnetogram-based (photospheric) features with coronal—thermal and morphological—features for flare forecasting. Furthermore, limiting the input to three EUV channels —whose observed structures are morphologically correlated— aims at leveraging optimally transfer learning from the use of CNN pretrained on standard red-green-blue (RGB) images, which also typically involve three morphologically correlated channels.

Flare windows labels. The labels are defined from the activity of the 24 hr time window, which starts from the sample date (i.e., the date at which we forecast the activity for the next 24 hr). Those binary flare labels (C+ or M+) are computed from an extended version (N. Plutino et al. 2024) of the N. Plutino et al. (2023) catalog, which uses GOES SXR flux data (H. A. Garcia 1994). To evaluate regional predictive skill on the operational test set, we also extract flare positions above the C-class threshold for events occurring after 2020 January 1. To that extent, we estimate flare coordinates on the solar disk by cross-referencing each event with 171 Å brightenings from SDO/AIA. Subtracting an image at the flare's onset from one at its peak isolates the dynamic intensity enhancement, which is then located via the `Trackpy` algorithm (J. C. Crocker & D. G. Grier 1996). This approach yields the event position for each flare above the C-class threshold, covering the period from 2020 January 1 onward.

Magnetograms. LOS magnetograms come from the HMI's 45 s SDO/HMI series archived by JSOC.[7] Each image is downsampled to 1024 × 1024 resolution (linear interpolation) and reduced from 16 bit to 8 bit depth. To retain a wide dynamic range without excessive saturation, we apply a symmetric log transform (i.e., $x \mapsto \mathrm{sign}(x)\ln(1 + |x|)$), then clamp pixel values above and below the 99.9th percentile of the magnetic-flux computed over the whole CV period (leading to a saturation value of ±4644 G). Finally, we linearly remap these values so that 0 and 255 correspond to negative and positive saturation, with the original zero field centered at 127.

EUV images. For coronal observations, we use the JSOC AIA synoptic data set[8] of 2 minute cadence, level 1.5 AIA images, already scaled and oriented so that solar north is up and aligned with the image's *Y*-axis. They are already downscaled from AIA images' native resolution to 1024 × 1024. Analogously to the magnetograms, we:

1. Reduce each image from 16 bit to 8 bit.
2. Normalize by exposure time and correct for CCD degradation over time using `aiapy` (W. T. Barnes et al. 2020).
3. Apply a log transform to preserve typical coronal signal (low and medium pixel values) and extreme flare brightenings (extreme pixel values) in a compressed range of values.
4. Saturate at the 99.9th percentile computed over the whole CV period and linearly rescale the data into $[0, 255]$.

We merge the 94 Å, 193 Å, and 211 Å channels into a single 8-bit three-channel image for use with standard pretrained

[7] JSOC series: hmi.M_45s.
[8] JSOC AIA synoptic data set http://jsoc.stanford.edu/data/aia/synoptic/

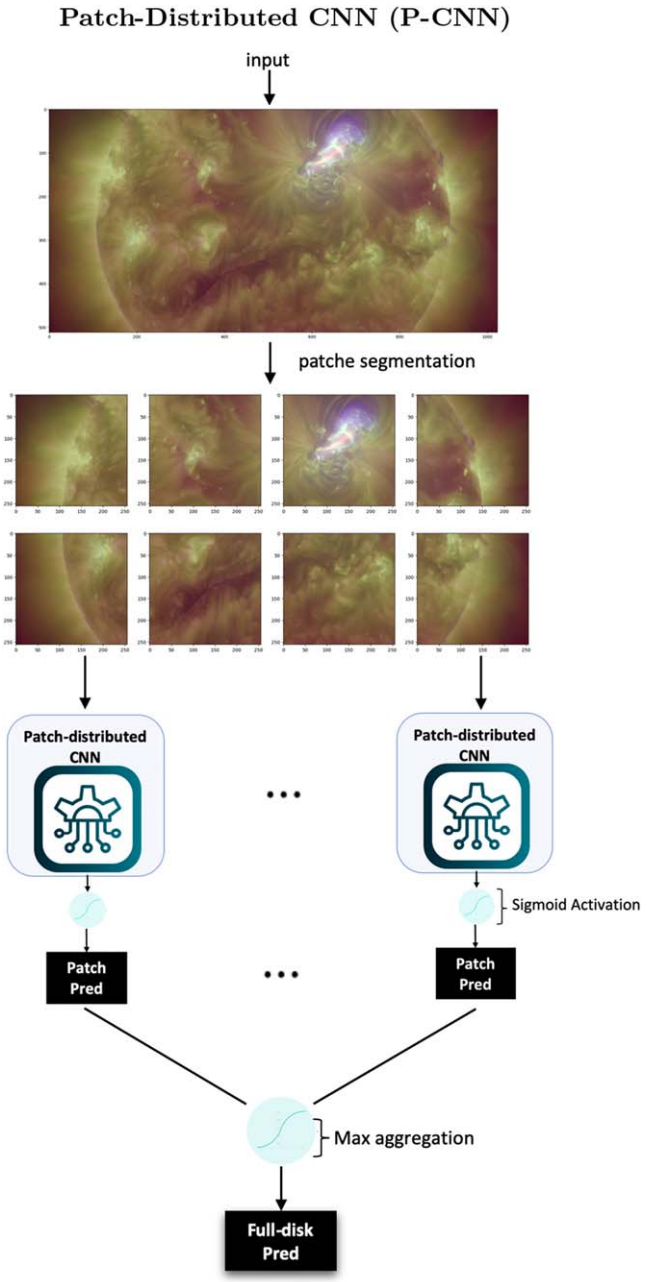

Figure 2. Architecture of the P-CNN. The input is segmented into patches, each processed by a CNN that applies identical weights and outputs sigmoids representing the flaring probabilities for the corresponding regions. The patch probabilities are max-aggregated to produce the model's final output representing the flaring probability at the whole disk level, which is the target directly learned during training. In this figure, the input combines 193 Å, 211 Å, and 94 Å EUV SDO/AIA images cropped at $\pm 614''$ downsampled at 224×448 pixels. The patches' size is set to 112×112 pixels, resulting in eight patches. In this work, an EfficientNetV2-S is used as the inner-patch-CNN.

CNN architectures. Additionally, single-wavelength 1600 Å, 304 Å, and 171 Å images are processed in a similar fashion, although not used in this work. All resulting 1024×1024 images of magnetograms and EUV channels are made publicly available as the SDO-2H-ML data set on Zenodo.[9]

Corrupted samples. After degradation correction, we identify corrupted images by locating observations whose mean pixel value lies outside an 8σ interval centered around a 48 hr running average. This strategy reveals 20–65 suspicious samples, depending on the channel, mostly caused by partial disk coverage (e.g., satellite maneuvers or eclipses). A few of these are valid but extreme events, especially in the 94 Å channel (very bright flares). We exclude problematic images (e.g., incomplete disks) from training and validation to avoid corrupted inputs, but leave them in the published data set with associated documentation.

3.2. Model: Patch-distributed CNN

To produce full-disk flare forecasts while still providing regional probability estimates, we segment full-disk inputs into non-overlapping patches. A unique CNN is distributed over each patch, outputting a sigmoid probability for the corresponding region. The final full-disk prediction is the maximum of the resulting patch's flare probabilities: $P_{\text{fulldisk}} = \max(\{P_{\text{patch}_i}\})$. We refer to this architecture as a patch-distributed CNN (P-CNN). The P-CNN model follows a weakly supervised deep learning approach, specifically employing inexact supervision, where full-disk labels are used to guide learning while the model internally generates patch-level predictions. This method enables the model to provide regional probability estimates despite the absence of patch-specific labels during training. The full-disk prediction is obtained through multiple instance learning aggregation, where the highest patch-level probability determines the final forecast. This strategy ensures that if any patch exhibits a high probability of flare occurrence, the full-disk prediction reflects this likelihood. Since flares predominantly occur at low solar latitudes, we crop poles above $\pm 614''$ to reduce input size by half and accelerate training. In our experiments, each cropped image is further downsampled to 224×448. We use 112×112 patches, yielding eight patches per image (Figure 2).

For the core (or inner-patch) CNN, we employ an EfficientNetV2-S (M. Tan & Q. V. Le 2021) initialized with ImageNet (J. Deng et al. 2009) pretrained weights. This model is then fully retrained (fine-tuned) to learn full-disk labels while internally generating patch-level predictions. We denote the magnetogram-based model variants as C+ magnetogram and M+ magnetogram; similarly, the EUV-based models are labeled C+ coronal and M+ coronal models, depending on whether they forecast flares above the C-class or M-class threshold.

3.3. Training and Evaluation

3.3.1. Full-disk CV Method

Following T. Cinto et al. (2020), we perform a chronological test split and keep all samples from 2020 January onward as an "operational" test set—free of any artificial sampling—while building a k-fold CV (where $k = 5$) on data from 2010 May to 2019 December.

Temporal chunking. Unlike AR-based modeling, where CV splits can be formed by separating unique AR numbers, full-disk forecasting presents strong temporal autocorrelations. We adopt a chunking strategy similar to E. J. E. Brown et al. (2022), grouping data into 81 day chunks separated by 27 day buffers (a full Carrington rotation), ensuring that each chunk is independent of others. Roughly 25% of the data is discarded by these buffer zones to maintain strong independence among CV folds. We allocate the resulting chunks across five folds in a balanced way, aiming for similar distributions of quiet, B, C,

[9] SDO-2H-ML data set doi:10.5281/zenodo.10465436 (G. Francisco et al. 2024).

M, and X flare classes, to foster optimal representation on each sub-case of positive and negative and mitigate the model's sensitivity to a specific climatology, as similarly suggested by (Z. Sun et al. 2022).

Balancing algorithm. Because chunks are large and some flare classes are rare, a perfect class match across folds is typically impossible without undersampling. To maximize balance before resorting to such undersampling, we employ an iterative algorithm that assigns each chunk to the fold that yields the greatest reduction in a "balance score" that we define with Equation (2):

$$b_k = \sum_{\mathrm{cls}\in\{\mathrm{quiet,B,C,M,X}\}} \delta_{\mathrm{cls}} \left| \frac{n^k_{\mathrm{cls}}}{\frac{N_{\mathrm{cls}}}{K}} - 1 \right|, \quad (2)$$

where δ_{cls} is the importance weight given to the balance of the class cls. Such a quantity is introduced to account for the impossibility of achieving a perfect balance. We set δ_{cls} to 4 for X, and for quiet-labeled time windows, we set it to 2 for M-labeled time windows, and we set it to 1 for B- and C-labeled time windows. In particular, it enables prioritizing achieving equal representation of the rarest classes before considering undersampling, as we aim to prevent further scarcity or the rarest events. n^k_{cls} is the number of time windows labeled as cls in the fold k. $\frac{N_{\mathrm{cls}}}{K}$ is the targeted number of samples of class cls for every fold, i.e., the ratio of N_{cls}, the whole number of time windows labeled as cls within the data set, with K the number of folds to be built.

Figure 3 illustrates the chunk allocation procedure, whereas Figure 4 depicts the post-balancing class counts.

Finally, we apply limited undersampling to produce:

1. *Training folds* with an approximately even composition of quiet, B, C, and M time windows, retaining all X-class samples.
2. *Validation folds* whose composition replicates the natural climatology of the whole CV period (roughly solar cycle 24).

By training on more balanced subsets, the model gains robust coverage of each subclass, while the validation sets will reflect realistic operational proportions (Figure 5).

3.3.2. Model's Hyperparameters

We implement our models with TensorFlow (M. Abadi et al. 2015), training on an NVIDIA V100 GPU with the following settings:

1. Optimizer: `Adam` (D. P. Kingma & J. Ba 2014) with decoupled weight decay (I. Loshchilov & F. Hutter 2017).
2. Learning rate: 10^{-5}; weight decay: 10^{-4}.
3. Batch size: 16 images.
4. Epochs: 15.

We save each fold's best model (selected by TSS on the validation set) and average them into an ensemble (average probability output) for final testing.

Loss function. We use weighted binary cross-entropy with the following training weights:

1. For C+ models, positives and negatives are equally weighted.

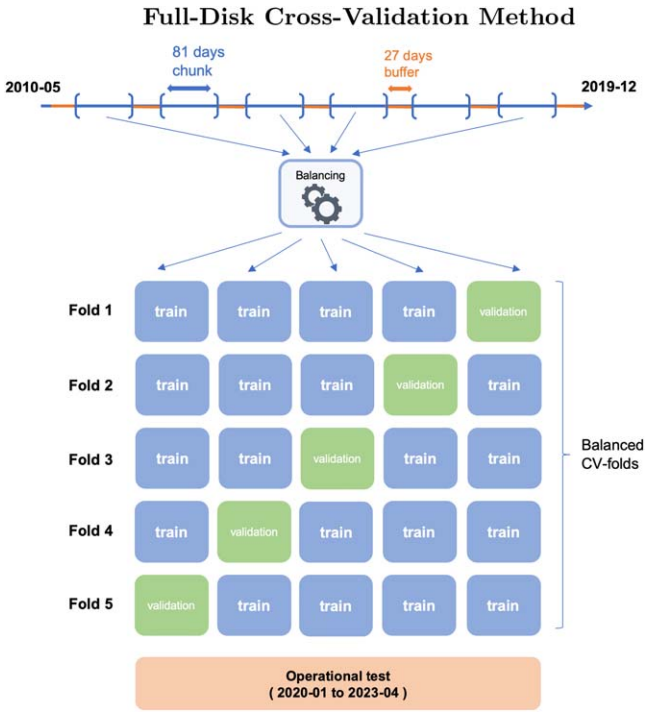

Figure 3. Temporal 81 day chunks separated by 27 day buffers (gray) form independent data segments for building training (blue) and validation (green) CV folds from 2010 May to 2019 December. All samples from 2020 January to 2023 April (orange) remain untouched as a fully chronological test set.

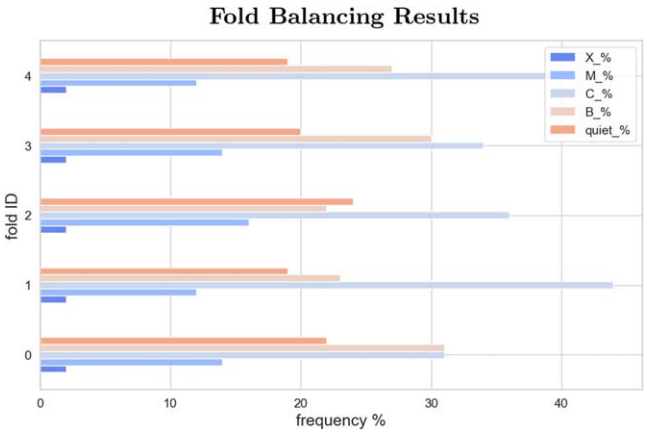

Figure 4. Folds' composition resulting from the pre-undersampling balancing algorithm (Section 3.3.1, paragraph—Balancing Algorithm) for 24 hr time windows.

2. For M+ models, we upweight positive events to improve recall, accepting a higher FAR as a trade-off. This approach highlights the limitations of TSS in imbalanced settings while aligning with the common objective of TSS maximization in the literature. Specifically, we assign weights of 2 to quiet and B-class time windows, 1 to C-class, and 8 to M- and X-class events. This weighting strategy encourages the optimizer to prioritize recall on rarer flares, illustrating how some metrics may fail to penalize models prone to high false alarm rates.

4. Results

This section presents and discusses the predictive performance of our models from different perspectives. Section 4.1 focuses on

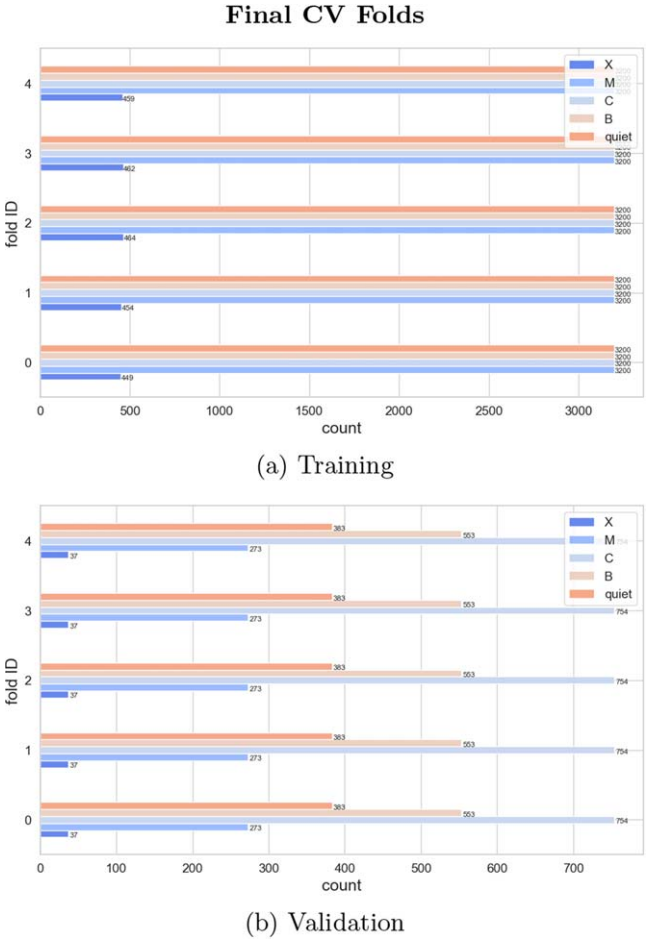

Figure 5. Final CV folds from undersampling of the balancing results in Figure 4. Validation replicates the CV period's climatology for better assessment of operational performance. Training's class even composition fosters balanced performances on each subclass and resilience to climatological variations.

full-disk results, noting both promising performances as well as the limiting aspects of conventional evaluations. Section 4.2 compares predictions at the regional (patch) level. Section 4.3 offers a brief visual explanation of model predictions, and Section 4.4 demonstrates how our method can retrieve sub-regional forecasts with a precise estimate of flare positions, thus providing an efficient weakly supervised learning framework.

4.1. Full-disk Performances

4.1.1. State-of-the-art Performances versus Low Persistence Relative Scores

Figures 6 and 7 display the distribution of TSS and HSS for the C+ and M+ models across validation and test sets, where the red dots indicate average single-fold performance, and the stars mark ensemble outcomes. Additional metrics, such as MCC and F1 score, appear in Table 1 for the test set. Overall, validation and test results align closely, indicating that our CV strategy successfully ensures independency between training and validation samples, thus avoiding the artificial increase of validation results through such dependencies coupled with overfitting. In fact, standard deviations of the fold-averaged metrics tend to be lower on the test set than on validation, suggesting that at least some observed variability during validation stems from residual data set biases in the folds.

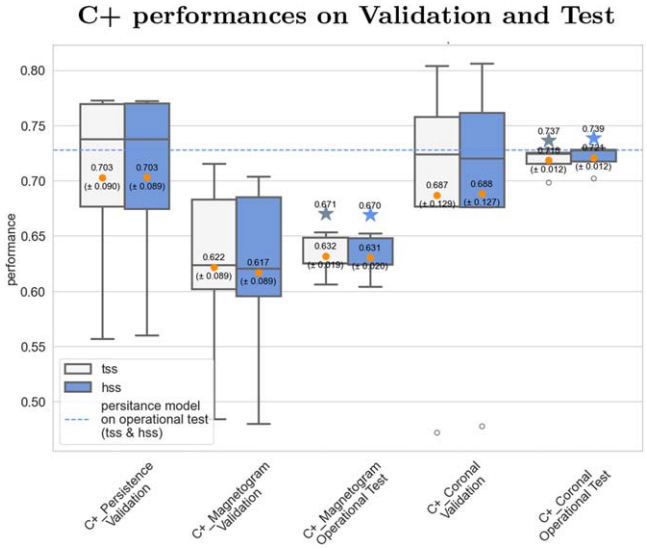

Figure 6. C+ models' performance. The box plots show TSS and HSS for each of the five CV-fold models. Red points denote fold-averaged results; the values beneath each red point indicate 1 standard deviation. The star symbols and labeled values above them represent the ensemble model's performance on the test set.

Coronal overperformance over magnetogram. For C+ forecasting, the magnetogram-based P-CNN achieves an average validation TSS of 0.62, while the coronal model attains 0.69. For M+ forecasting, the magnetogram and coronal models reach TSS values of 0.53 and 0.57, respectively. On the operational test set (2020 January–2023 April), we ensemble the five CV-fold models by averaging their predicted probabilities. These ensemble models outperform each individual fold's performance; in particular, the C+ coronal model reaches a TSS of 0.74, and the M+ coronal model 0.61.

The overperformance of models with coronal inputs appears to be systematic and verified across the several operational metrics listed in Table 1.

Comparisons to persistence and other works. Despite these apparently strong results, a closer look reveals two crucial caveats:

1. *Low persistence relative skill.* As Table 1 shows, a simple persistence approach—predicting that each time window will continue the same flare activity (or inactivity) as the previous window—achieves TSS scores exceeding 0.70 for C+ and around 0.45 for M+. More critically, metrics that incorporate precision, such as MCC, HSS, or F1 score, do not indicate a significant advantage for our models over the persistence model. In fact, the persistence relative F1 (PR-F1) column in Table 1 is always near zero, implying that, in terms of compromising between detection and false alarms, the deep learning models offer little improvement over the persistence model. For M+ in particular, boosting TSS comes largely from overcasting (a high recall but low precision), a known pitfall allowed by imbalanced data.
2. *Challenges in cross-study comparisons.* Because TSS, HSS, and even MCC are highly sensitive to data set composition (Section 2.1 and Appendix A.2.1), it is difficult to make direct comparisons with previous flare forecasting literature. Nonetheless, the HSS values of our

Table 1
Full-disk Performance Summary of the Operational Test Set

Models	TSS	HSS	MCC	F1	Recall	FAR	ϕ	χ	PR-F1	AC-MCC	NC-MCC	NC-ϕ
C+ persistence	0.73	0.73	0.73	0.86	0.86	0.14	0.48	0.14	0	−1	1	0.48
C+ coronal	0.74	0.74	0.74	0.86	0.82	0.10	0.48	0.14	−0.00	0.13	0.84	0.48
C+ magnetogram	0.67	0.67	0.68	0.82	0.77	0.11	0.51	0.14	−0.04	0.03	0.78	0.50
M+ persistence	0.45	0.45	0.45	0.53	0.53	0.47	0.14	0.13	0	−1	1	0.08
M+ coronal	0.61	0.42	0.46	0.53	0.82	0.61	0.14	0.13	−0.00	0.08	0.47	0.09
M+ magnetogram	0.58	0.37	0.43	0.50	0.84	0.65	0.14	0.14	−0.06	0.06	0.43	0.09

Note. Final ensemble models' results on the test set for both the C+ and M+ models. $\phi = \frac{P}{P+N}$ is the positive event ratio; $\chi = \frac{C}{P+N}$ is the AC rate (the fraction of windows where the label differs from the previous window). PR-F1 is the F1-based persistence relative skill (Equation (1)). AC-MCC and NC-MCC are MCC restricted to AC and NC windows, respectively. NC-ϕ is the positive event ratio specifically among NC windows.

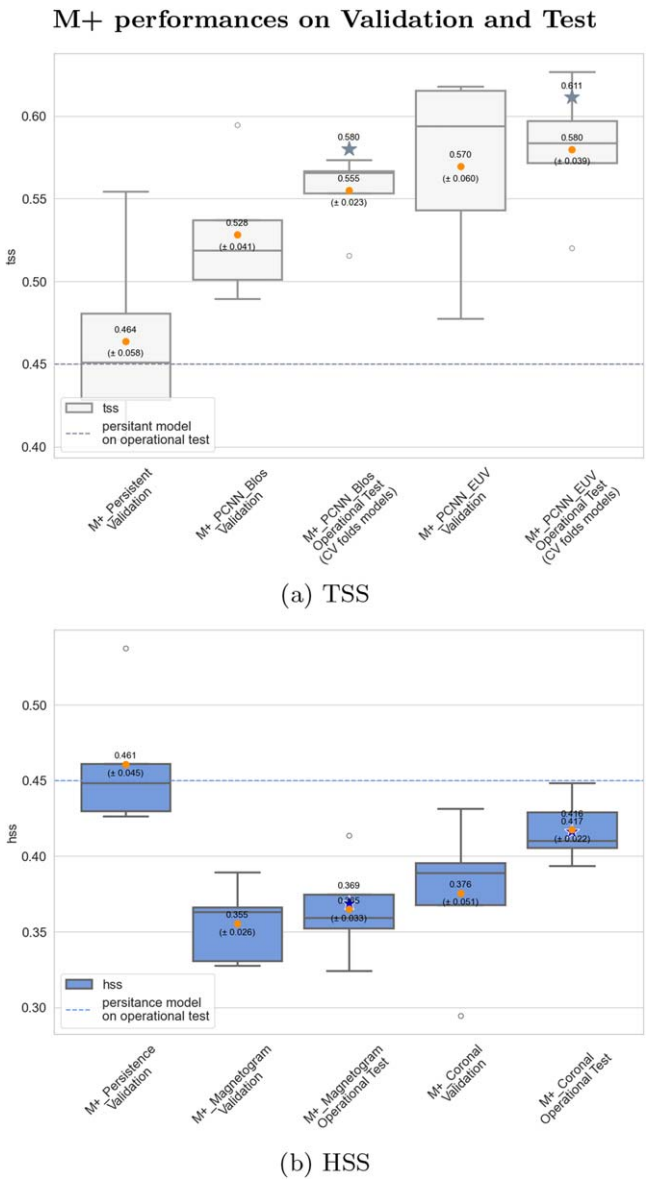

Figure 7. M+ models' performance. The box plots summarize TSS (top) and HSS (bottom) across the five CV-fold models. Red points and the numeric labels above each set represent average fold performance, with standard deviations beneath in smaller text. Star markers (with labels above) show the ensemble performance on the test set.

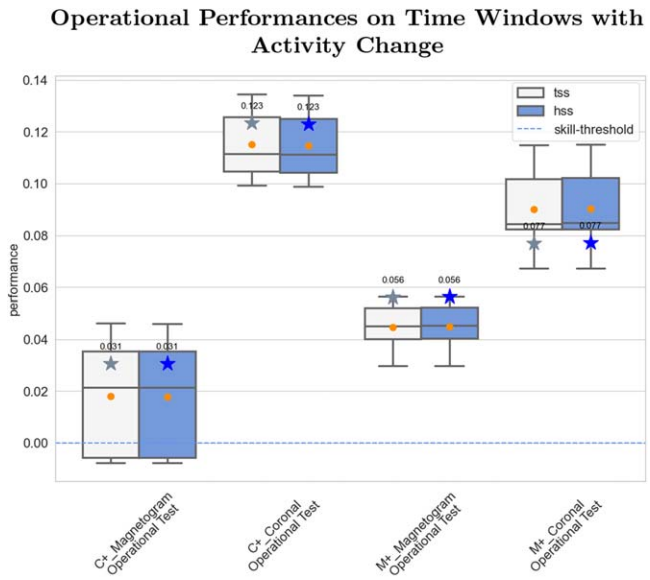

Figure 8. Models' operational metrics restricted to windows in which the activity label changes (AC windows). Box plots summarize each fold, red points indicate the average fold value, and star markers give the ensemble performance. All three metrics drop significantly compared to the full-disk results in Figures 6 and 7.

models compare favorably to other reports, indicating competitive or superior performance under this comprehensive metric.

4.1.2. Performance Deficiency on ACs

To isolate cases where the flare label differs from that of the previous time window, we use the AC metrics defined in Section 2.2. This consists of evaluating time windows transitioning from inactive to active, and those moving from active to inactive (first column examples in Figure 1). Figure 8 and Table 1 reveal notably low AC-HSS, AC-TSS, and AC-MCC, implying that models barely outperform random guessing on the subset of events where the activity is changing (first flare or first quiet period).

4.2. Regional Performances

Table 2 presents results at the patch (regional) level. These scores tend to be lower than their full-disk counterparts,

Table 2
Regional Performance Summary of the Operational Test Set

Models	TSS	HSS	MCC	F1	Recall	FAR	ϕ	χ	PR-F1	AC-MCC	NC-MCC	NC-ϕ
C+ persistence	0.57	0.57	0.57	0.65	0.65	0.35	0.18	0.13	0	−1	1	0.14
C+ Coronal	0.50	0.57	0.58	0.63	0.54	0.22	0.18	0.13	−0.02	0.04	0.73	0.14
C+ magnetogram	0.39	0.48	0.51	0.55	0.41	0.19	0.19	0.14	−0.16	0.04	0.66	0.14
M+ persistence	0.32	0.33	0.33	0.34	0.34	0.67	0.02	0.03	0	−1	1	0.01
M+ coronal	0.52	0.28	0.32	0.30	0.56	0.80	0.02	0.03	−0.12	0.06	0.29	0.01
M+ magnetogram	0.43	0.22	0.26	0.25	0.48	0.83	0.02	0.03	−0.27	0.09	0.20	0.01

Note. Columns follow the same definitions as Table 1. The patch-level data set is more imbalanced, causing a further drop in metrics.

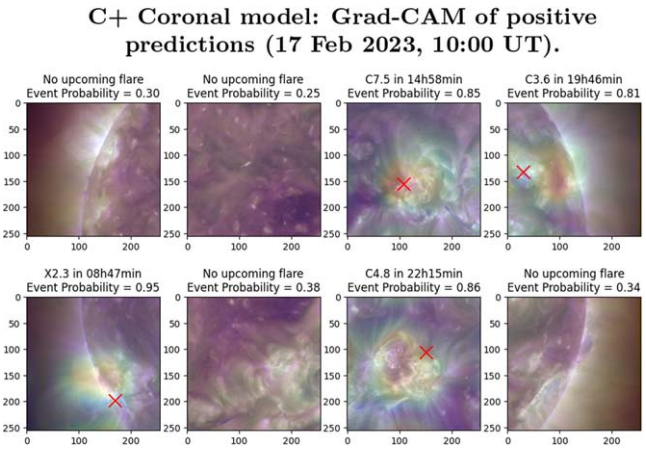

Figure 9. Grad-CAM explanation maps for the C+ coronal model's patch outputs. The ground truth (largest flare in the next 24 hr) appears in the first line of each patch title, while the second line shows the model's predicted class. Red-to-blue contours highlight key regions used by the CNN. Red crosses show the upcoming flare location (with solar rotation accounted for). The model's most discriminative patch areas coincide with ARs set to flare.

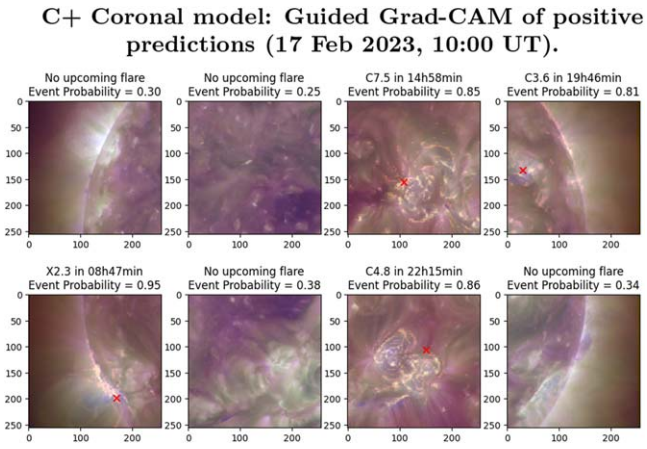

Figure 10. Guided Grad-CAM highlights the fine-scale pixels most influential for the patch-level classification. The "hot" color scale shows that intense coronal loops or bright emission dominate the model's reasoning.

reflecting less favorable data set composition. For instance, in the M+ case, the AC rate and positive event ratio fall below 3% at the regional level, leading to even lower skill scores for both the P-CNN and the baseline persistence models.

4.3. Explainability

Figures 9 and 10 illustrate an example explainability analysis for the C+ coronal model on 2023 February 17, at 10:00 UT, about 9 hr before a powerful X2.3 flare. Within the subsequent 24 hr, four regions (including the limb that produced the X2.3 flare) indeed hosted C-class or larger flares.

Grad-CAM analysis. To identify which regions most influenced each patch prediction, we apply Grad-CAM (R. R. Selvaraju et al. 2016). Grad-CAM heatmaps (Figure 9) show that the model focuses on the ARs that actually flare within the next 24 hr, implying that it successfully exploits relevant bright coronal features in each patch.

Guided Grad-CAM. Figure 10 further refines these maps by combining guided backpropagation (J. T. Springenberg et al. 2014) with Grad-CAM, yielding fine-grained feature localization. Bright coronal structures dominate the network's focus, consistent with the finding that the model essentially recognizes features that correlate with already ongoing high-energy rather than anticipating major changes in the near future.

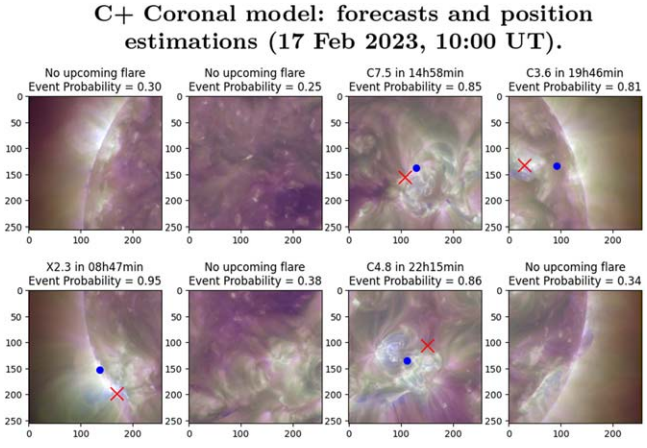

Figure 11. Example of patch-level probabilities and position estimates for the C+ coronal P-CNN models. The red cross denotes the true flare location after rotation correction, and the blue dot is the centroid of the Grad-CAM heatmap. All four patches hosting a future flare are correctly labeled as positive, with position estimates matching the associated ARs.

4.4. Positions Predictions

Finally, we estimate potential flare locations by computing the centroid of the Grad-CAM intensities within each positively classified patch. Such approach could then be used to cross-referenced eventually known ARs or to label new ARs not yet present in available databases. Figure 11 demonstrates how the inferred positions (blue dots) align closely with the actual

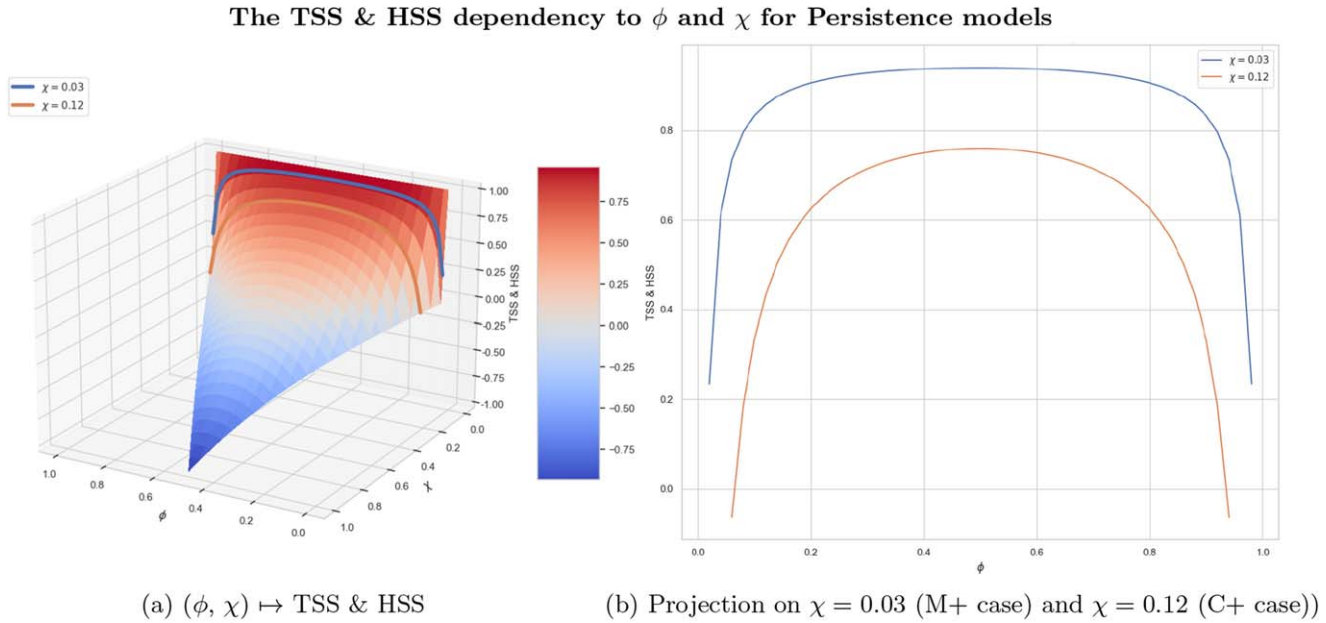

Figure 12. Figure (a) displays the plot of a persistence model's TSS, HSS, and MCC as a function of ϕ and χ. Figure (b) displays the variations of the TSS, HSS, and MCC with ϕ for two different constant χ. $\phi = \frac{P}{P+N}$ is the positive event ratio. $\chi = \frac{C}{P+N}$ is the AC rate.

emerging flares (red crosses) for the four events, each accurately predicted.

In summary, this approach offers promising operational value: identifying position estimates within disk regions most likely to produce a flare and overcoming potential errors from AR misidentification from an external system.

5. Discussion

5.1. P-CNN: An Efficient Weakly Supervised Approach to Operational Models

Greater information retrieval with P-CNNs. The P-CNN introduced here reconciles the simplicity of end-to-end full-disk training with the fine-grained regional insight typically associated with AR-based methods. While each sample is labeled only at the full-disk level, the patchwise approach enables per-region flare probability outputs and position estimations through Grad-CAM analyses. Operationally, this design can inform users not just of an upcoming flare but also of where to look on the solar disk. A potential caveat arises for AR spanning multiple patches. Although these cases are generally detected, the model may partially misattribute the flare signal to one patch over another. This is, however, partly mitigated with Grad-CAM centroids, which allow a more precise localization.

Potential regularization from P-CNNs. In practice, our best-performing EUV-based models trained with the P-CNN architecture proved simpler to tune and delivered stronger performance than a conventional full-disk CNN. We hypothesize that P-CNN enforces a form of regularization: each patch CNN must focus on local features (spatial scales that are naturally comparable to AR), potentially smoothing optimization and reducing overfitting to large-scale image artifacts. Additionally, because the model shares the CNN weights across all patches, each training sample effectively multiplies the number of relevant "sub-instances" by the number of patches, further increasing regularization. While we observed such advantages of P-CNN on EUV images, improvements were less pronounced for magnetogram inputs, which may stem from either reduced discriminative magnetogram features or suboptimal use of ImageNet pretrained weights for non-RGB data. A more exhaustive hyperparameter search would be needed to confirm these conclusions. Finally, to also benefit from global context (e.g., extended coronal loops or multi-AR interactions), one could augment a P-CNN with a global CNN branch, yielding a "pyramidal" or multi-scale feature approach.

5.2. Evaluating Model Performance with PR-F1

In this work, we introduce the PRSS metrics to provide a more informative assessment of model performance in solar flare forecasting. While traditional skill metrics such as TSS and HSS are widely used, they are also highly sensitive to data set composition and can hide strong operational limitations when evaluated on imbalanced data sets. In contrast, PRSS metrics normalize the model performance relative to a persistence-based baseline, offering a complementary reference point for comparison. Among the proposed PRSS metrics, we emphasize PR-F1 as the most practical and reliable for performance evaluation. PR-F1 integrates both precision and recall for the positive class, making it particularly suited for imbalanced classification tasks, where TSS alone may be misleading. PR-F1 exhibits greater stability across different data set compositions, maintaining consistency even when the test set is sampled at different phases of the solar cycle. As shown in Annex A, Figure 12, metrics like the TSS are very sensitive to dataset compositions biases as the positive event ratio and the AC rate. Empirically however, the PR-F1 stands out as the most robust to variations of those biases, as displayed by Figures 13 and 14 in Annex B. As such, the PR-F1 can offer better consistency to evaluate on test sets sampled at different phases of the Solar Cycle. While PR-TSS and PR-HSS can still offer useful insights by quantifying the improvement over persistence-based models in relative terms, their double normalization introduces interpretability challenges, making them less suitable as primary evaluation metrics.

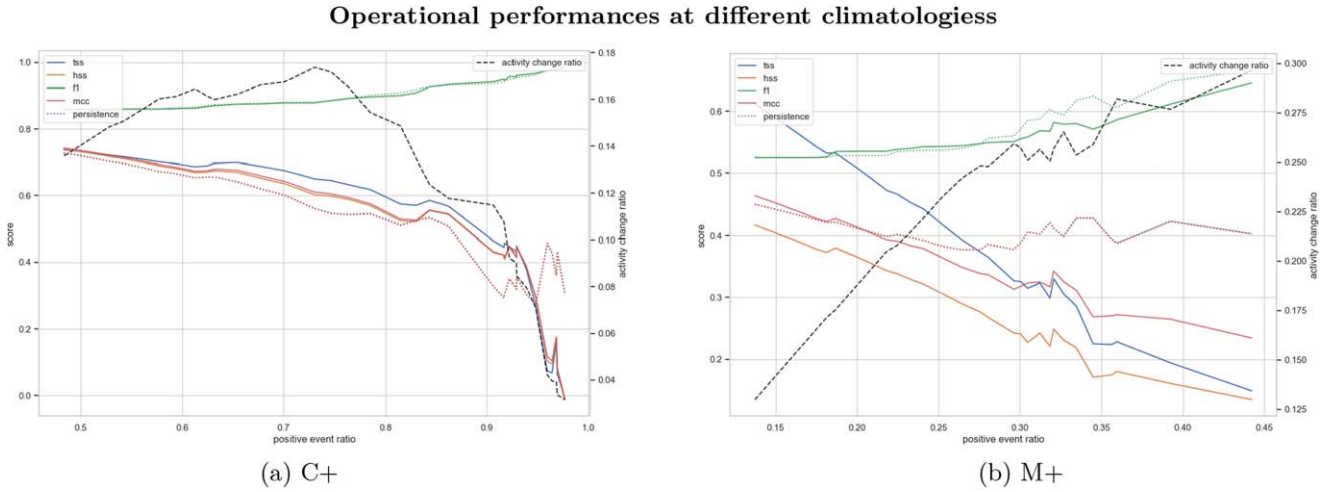

Figure 13. EUV models' operational performances against variations of the positive event ratio and the AC rate. The sub-test sets with varying compositions are obtained by sliding the start of the test set from 2020 January 1 to 2023 January 1. The left *Y*-axis represents the metrics score. The *X*-axis is the varying positive event ratios. The right *Y*-axis represents the AC rates, which are plotted as black dashed lines.

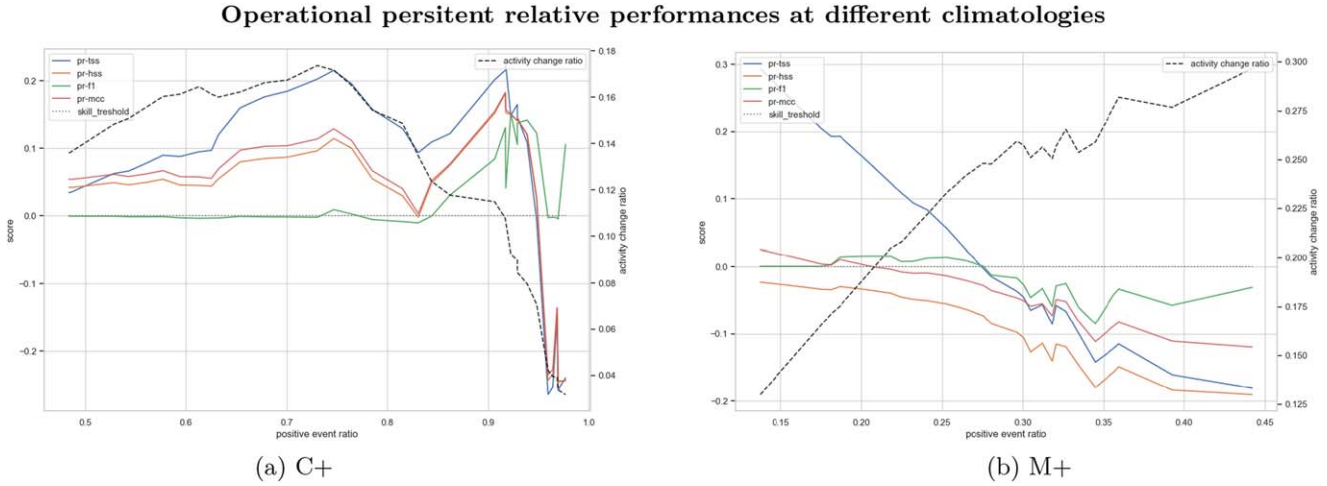

Figure 14. EUV models' persistence relative performances against variations of the positive event ratio and the AC rate. The axis and sub-test sets are the same as in Figure 13.

5.3. Current Models' Weaknesses

Our analysis reveals that even state-of-the-art TSS and HSS values may hide significant weaknesses. The PR-F1 scores near zero indicate that the models' precision and recall collectively do not surpass those of the persistence model, translating into limited abilities for a relevant alert system. The weak AC-MCC values confirm the models' struggle to predict genuine changes in flaring behavior better than random guesses. These findings suggest that the features currently learned from photospheric magnetograms and coronal EUV emissions primarily help to recognize and classify ongoing activity and ongoing quiet configurations, rather than anticipating changes between these two states.

Potential directions to address this limitation include:

1. Incorporating time series models: Z. Sun et al. (2022) reported a 5%–11% gain by ensembling CNN (image-based) with LSTM (time series features). Evaluating AC performance of such hybrid methods might determine whether they can indeed capture dynamical changes significantly better than random chance.
2. Exploring alternative flare labels: Instead of classifying flares strictly by maximum flux (MPF), one could forecast the integrated flux (SXR fluence) over a time window. This could align more closely with magnetogram-based proxies (e.g., C. J. Schrijver (2009)'s R-index) that correlate with unstable stored magnetic energy.
3. Investigating additional multimodal temporal features. Our preliminary attempts (e.g., including flare history or SXR flux time series with LSTM) did not outperform the persistence baseline, suggesting that a Markov-like memory is possibly already embedded in the magnetogram/EUV representations at a given instant.

Finally, we note that the resolution and compression of our images (JPEG 224 × 448 downsampled crops) had a negligible impact on performance over doubled spatial resolution and lossless compression, suggesting that small-scale nontemporal features—lost through downsampling and JPEG compression

—may offer limited further predictive information for 24 hr forecasts.

5.4. AC Definition Limit

Although distinguishing time windows with AC from those with NC elucidates important limitations, this definition has inherent constraints. For instance, if an AR emerges from the farside of the Sun, the visible full-disk could transition from "inactive" to "active" and be labeled as AC even though the corresponding AR could have been already flaring on the farside, thus corresponding to a constantly flaring region which corresponds to an NC window. Similarly, AR movement from one patch to another can mimic an AC at the patch level, when physically it is the same AR with continuous activity. Additionally, the 24 hr window can artificially label time windows with flares spaced 23 hr apart as "constantly active (NC)" but flares spaced 25 hr apart as consecutive windows presenting "change of activity (AC)" These artifacts generally inflate the apparent AC performance since some "AC windows" physically correspond to near-constant activity. More refined temporal metrics, such as those proposed by S. Guastavino et al. (2022), could mitigate these issues by weighting errors according to their temporal offset. Nonetheless, the core finding stands: models using point-in-time (or one time step) features and displaying state-of-the-art performances (in HSS and TSS), do not reliably forecast first flares and first quiet periods, a crucial challenge for future research.

6. Conclusion

Our study introduces a novel method to facilitate the construction of balanced and independent CV folds for full-disk flare forecasting with minimal undersampling. Introducing P-CNNs trained with such a method, our models achieve state-of-the-art performances at full-disk levels on the rising phase of solar cycle 25, with a TSS of 0.74 for C+ forecasts and 0.62 for M+ forecasts using EUV coronal images as inputs, demonstrating a consistent advantage over magnetogram-based predictions. The P-CNN requires only full-disk labels yet provides subregional (patch-level) forecasts with position estimates, and no reliance on external AR-detection or labeling. This results in a weakly supervised learning framework that mitigates common issues of misattributed or missing flares from AR-level flare catalogs, thereby simplifying training and evaluation. Our findings also underscore shortcomings of common metrics in flare forecasting, given their high sensitivity to data set composition and inability to highlight critical model weaknesses, particularly in predicting ACs. To address these concerns, we introduce (1) PRSSs (e.g., PR-F1) to benchmark models against a competitive no-skill persistence, and (2) restricted evaluations on time windows with and without changes of activity with respect to the previous period (AC and NC windows). These additional metrics, tailored to account for the imbalanced and dynamic nature of flare events, help identify and assess forecasting strengths and weaknesses. In particular, PR-F1 suggests that our models' performances barely exceed the persistence ability to accurately predict an event with a low FAR, and the AC-MCC highlights the models' low skill in forecasting changes in flaring behavior—performing marginally better than random classifiers on such periods (AC windows). Finally, both the PR-F1 and MCC exhibit greater stability and reliability than the HSS, the latter having already been shown to be more reliable and informative than the TSS in imbalanced scenarios. Overall, our results motivate future research toward multimodal temporal features and architectures specifically aimed at capturing AR emergent and changing activity.

Additionally, we plan to explore optimizations to the P-CNN model architecture by incorporating region-specific CNNs to account for varying projection effects across different areas of the solar disk. Currently, the P-CNN model shares weights across all patches, requiring a single model to generalize across different regions of the solar disk despite significant differences in viewing angles and projection effects. This uniform weight-sharing scheme does not leverage the unique spatial properties of each patch, which may put unnecessary constraints on the model's learning capabilities. One promising avenue for improvement is adopting a mixture-of-experts approach (N. Shazeer et al. 2017), where separate CNNs specialize in different regions of the solar disk, allowing each model to better capture region-specific features. For instance, patches closer to the solar center provide an unobstructed view of ARs, whereas patches at the limb suffer from foreshortening and projection effects. A dedicated CNN for limb patches could learn projection-invariant representations, while central patches could focus on extracting fine-grained spatial details. This specialization could enhance predictive performance by reducing the burden on a single CNN to generalize across all regions. To mitigate the computational challenges of region-specific CNNs, future work will explore parameter-efficient strategies that enable specialization while maintaining efficiency. Approaches such as adaptive fine-tuning, shared representations, or multi-task learning could enhance model adaptability without significantly increasing complexity (I. Kokkinos 2017; S. Ruder 2017; A. Aghajanyan et al. 2020; E. J. Hu et al. 2021). Evaluating these techniques will help balance performance improvements with practical constraints in operational forecasting.

7. Software and Third-party Data Repository Citations

The data prepared for this study offers a compact and machine learning ready data set that can be used for other applications and is available as the SDO-2H-ML data set on Zenodo.[10] It is derived from the AIA synoptic data set[11] and JSOC's level 1.545 s series HMI LOS magnetograms. The time window labels are derived from an extension (N. Plutino et al. 2024) of the Plutino flare event catalog (N. Plutino et al. 2023).

The code based on `TensorFlow` (M. Abadi et al. 2015) was developed to analyze and train the models of this work, and notebooks to replicate the results presented in this study are available on the Zenodo repository[12] (G. Francisco 2025).

Acknowledgments

This research is part of the SWATNet project, which is funded by the European Union's Horizon 2020 research and innovation program under the Marie Sklodowska-Curie grant Agreement No. 955620.

This research has also been carried out in the framework of the CAESAR project, supported by the Italian Space Agency and the National Institute of Astrophysics through the ASI-INAF n.2020-

[10] SDO-2H-ML: doi:10.5281/zenodo.10465436 (G. Francisco et al. 2024)
[11] AIA synoptic data set: http://jsoc.stanford.edu/data/aia/synoptic/
[12] doi:10.5281/zenodo.14790146

35-HH.0 agreement for the development of the ASPIS prototype of the scientific data center for Space Weather.

This study was also produced within the IA and the CITEUC. IA is supported by Fundação para a Ciência e a Tecnologia (FCT, Portugal) through the research grants UIDB/04434/2020 and UIDP/04434/2020. CITEUC, is funded by National Funds through FCT—project UIDP+UIDB/00611/2019.

Michele Berretti acknowledges that this publication (communication/thesis/article, etc.) was produced while attending the PhD program in Space Science and Technology at the University of Trento, Cycle XXXIX, with the support of a scholarship financed by the Ministerial Decree no. 118 of 2023 March 2nd, based on the NRRP—funded by the European Union—NextGenerationEU - Mission 4 "Education and Research," Component 1 "Enhancement of the offer of educational services: from nurseries to universities"— Investment 4.1 "Extension of the number of research doctorates and innovative doctorates for public administration and cultural heritage"

Project partially funded under the National Recovery and Resilience Plan (PNRR), Missione 4 "Istruzione e Ricerca"—Componente C2—Investimento 1.1, "Fondo per il Programma Nazionale di Ricerca e Progetti di Rilevante Interesse Nazionale (PRIN)"— Call for tender No. 1409 of 14/09/2022 of Italian Ministry of University and Research funded by the European Union—NextGenerationEU Award No.: P2022RKXH9, Concession Decree No. 1397 of 06/09/2023 adopted by the Italian Ministry of University and Research, project CORonal mass ejection, solar eNERgetic particle and flare forecaSTing from phOtospheric sigNaturEs (CORNERSTONE).

List of Acronyms

1D	One-dimensional
2D	Two-dimensional
3D	Three-dimensional
AA	Academy of Athens
AC	Activity change
AIA	Atmospheric Imaging Assembly
ASRO	Aboa Space Research Oy
AI	Artificial intelligence
AR	Active region
ASCII	American Standard Code for Information Interchange
AU	Astronomical unit
CAM	Class Activation Maps
CDAWeb	Coordinated Data Analysis Web
CDF	Common Data Format
CDP	Career development plan
CET	Central European Time
CME	Coronal mass ejection
CNN	Convolutional neural network
CSV	Comma-separated values
CV	Cross validation
DHO	Debrecen Heliophysical Observatory
DOI	Digital object identifier
EAB	External Advisory Board
ECAS	European Commission Authentication System
ECTS	The European Credit Transfer and Accumulation System
ELTE	Eötvös Loránd University
ESA	European Space Agency
ESR	Early-stage researcher
EU	European Union
EUV	Extreme ultraviolet
EUHFORIA	European Heliospheric Forecasting Information Asset
EUHFORIA 2.0	European Heliospheric Forecasting Information Asset 2.0
FAR	False alarm ratio
FGE	Fluid gravity
FITS	Flexible Image Transport System
FN	False negative
FP	False positive
FP7	The Seventh Framework Programme of the European Union
FSS	F1 skill score
GA	Grant agreement
GAN	Generative adversarial network
GCS	Graduated cylindrical shell
GSO	Gyula Bay Zoltán Solar Observatory
HEEQ	Heliocentric Earth equatorial
HMI	Helioseismic and Magnetic Imager
HSPF	Hungarian Solar Physics Foundation
HSS	Heidke skill score
ICME	Interplanetary coronal mass ejection
IDL	Interactive Data Language
I/O	Input/output
ITN	Innovative training network
IMF	Interplanetary magnetic field
IP	Interplanetary
IPN	Instituto Pedro Nunes
JSOC	Joint Science Operations Center
KUL	KU Leuven
L1	First Lagrangian point
LOS	Line of sight
LSTM	Long short-term memory
MCC	Matthews correlation coefficient
MSCA	Marie-Skłodowska-Curie action
MFM	Magnetofrictional method
MHD	magnetohydrodynamics
ML	Machine learning
MLP	Multi-layer perceptron
MPF	Maximum peak flux
NC	No change
NFB	Negative frequency bias
NN	Neural network
NPV	Negative predictive value
NRT	Near real time
OA	Open access
P-CNN	Patch-distributed CNN
PHI	Polarimetric and Helioseismic Imager
PIL	Polarity inversion line
PCDP	Personal Career Development Plan
PFSS	Potential field source surface
PI	Principal Investigator
PPV	Positive predictive value
PR-F1	Persistence relative F1
PRSS	Persistence relative skill score
PSP	Parker Solar Probe
PTECH	Present Technologies, LDA
RGB	Red-green-blue
RNN	Recurrent neural network
ROC AUC	Receiver Operating Characteristic's Area Unde the Curve
SAS	Space Applications Services
SB	Supervisory Board
SDO	Solar Dynamic Observatory
SEP	Solar energetic particle
SF	Solution focus
SOHO	Solar and Heliospheric Observatory
SolO	Solar Orbiter

(Continued)

SOLPACS	Solar Particle Acceleration in Coronal Shocks
SPCNN	Solar patch-distributed CNN
SSC	Sheffield Solar Catalog
STEM	Science, Technology, Engineering, and Mathematics
STEREO	Solar Terrestrial Relations Observatory
SVM	Support vector machine
SWATNet	Space Weather Awareness Training Network
SXR	Soft X-Rays
TN	True negative
TP	True positive
TPR	True positive rate
TSS	True skill statistic
WP	Work package
UC	University of Coimbra
UH	University of Helsinki
UMCS	Maria Curie-Skłodowska University
UNITOV	Universitá degli Studi di Roma Tor Vergata
UoI	University of Ioannina
USFD	University of Sheffield
UTU	University of Turku
UV	Ultraviolet

Appendix A
Metrics Complements

A.1. Formulas

A.1.1. Confusion Matrix

Evaluation metrics for binary classification are defined as a function of the confusion matrix (CM):

$$\mathrm{CM} = \begin{pmatrix} \mathrm{TP} & \mathrm{FN} \\ \mathrm{FP} & \mathrm{TN} \end{pmatrix}, \tag{A1}$$

where TP and TN denote the number of positive and negative events correctly classified, while FP and FN are the number of misclassified ones.

A.1.2. Basic CM Rates

The following rates summarize the information contained in the CM:

1. Class accuracy rates

$$\mathrm{recall} = \mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{A2}$$

$$\mathrm{sensitivity} = \mathrm{TNR} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}} \tag{A3}$$

2. Class precision rates:

$$\mathrm{NPV} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FN}}, \tag{A4}$$

$$\mathrm{precision} = \mathrm{PPV} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}. \tag{A5}$$

Equivalently to the precision, practitioners alternatively look at the FAR, which is the complementary of the former and gives the rate at which positive predictions give a false alarm.

A.1.3. F1 Score

The F1 score, defined in Equation (A6), is the harmonic mean of the precision and the recall.

$$\mathrm{F1} = 2 * \frac{\mathrm{precision} * \mathrm{recall}}{\mathrm{precision} + \mathrm{recall}} = \frac{2\mathrm{TP}}{2\mathrm{TP} + \mathrm{FP} + \mathrm{FN}} \tag{A6}$$

It offers a consolidated assessment of an alarm system skill when the emphasis lies on achieving high recall (TPR) to detect maximum events, along with high precision (PPV) to ensure confidence in positive predictions.

A.1.4. TSS/Informedness

The TSS was introduced to evaluate weather forecasts by (A. Hanssen & W. Kuipers 1965). In other fields, it is also known as the (bookmaker) informedness, Peirce's index, or Younden's J index. It can be dated back to 1884 (C. S. Peirce 1884). It is equal to the difference between the TPR and the false positive rate but also to the balanced accuracy re-scaled between −1 and 1, i.e., the average of the class accuracy rates (TPR and TNR) normalized in [−1,1]:

$$\mathrm{TSS} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} - \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}} = \mathrm{TPR} + \mathrm{TNR} - 1. \tag{A7}$$

Random and constant models produce a TSS of 0.

A.1.5. Markedness

The markedness (MK) is the precision equivalent of the TSS; it is the average of the class precision rates (PPV and NPV) normalized in [−1,1].

$$\mathrm{MK} = \mathrm{PPV} + \mathrm{NPV} - 1. \tag{A8}$$

A.1.6. Negative Frequency Bias

$$\mathrm{NFB} = \frac{\mathrm{TN} + \mathrm{FN}}{\mathrm{TN} + \mathrm{FP}} = \frac{\text{predicted negatives}}{\text{observed negatives}}. \tag{A9}$$

A.1.7. HSS/Cohen's Kappa Index

The HSS, defined in Equation (A10), was introduced to evaluate weather forecasts by P. Heidke (1926). In other fields, it can be known as Cohen's kappa index. It is commonly used in flare forecasting to compare a model's skill relatively to a random guess model (E. Camporeale 2019).

$$\mathrm{HSS} = 2 * \frac{\mathrm{TP} * \mathrm{TN} - \mathrm{FN} * \mathrm{FP}}{\mathrm{P}(\mathrm{TN} + \mathrm{FN}) + \mathrm{N}(\mathrm{TP} + \mathrm{FP})} \tag{A10}$$

The HSS varies between −1 and 1, with 1 denoting the performance of a perfect classifier and 0 indicating the one of random guesses. It can then be noted that the HSS is the harmonic mean between $\frac{\mathrm{TSS}}{\mathrm{NFB}}$ and MK * NFB (see (R. Delgado & X.-A. Tibau 2019 for mathematical proof). The HSS is, therefore, a weighted harmonic average between the TSS and the MK, with a model-dependent importance given to each.

A.1.8. MCC

The MCC was introduced by B. Matthews (1975) to address class imbalances in performance evaluation. It is the Pearson correlation coefficient between binary predictions and labels:

$$\mathrm{MCC} = \frac{\mathrm{TP} * \mathrm{TN} - \mathrm{FN} * \mathrm{FP}}{\sqrt{\mathrm{P}(\mathrm{TN} + \mathrm{FN}) * \mathrm{N} * (\mathrm{TP} + \mathrm{FP})}}. \quad \text{(A11)}$$

The MCC ranges between -1 and 1. Similar to the TSS and HSS, both random and constant models produce an MCC score of 0. The MCC is the geometric average between the TSS and the markedness (see R. Delgado & X.-A. Tibau 2019; D. Chicco et al. 2021a). It thus summarizes the four basic confusion matrix rates with equal weights given to each.

A.2. Metrics' Notable Properties

A.2.1. The TSS Sensitivity to the Data Set Composition

D. S. Bloomfield et al. (2012) proposed the TSS for flare forecasting, as in simplified cases, it is found to be insensitive to the class balance. This has been argued to make it a suitable metric for comparing models among different data sets with varying class balances (D. S. Bloomfield et al. 2012; D. Chicco et al. 2021a). However, in the case of flare forecasting, we prove that usual models will have a TSS that is strongly sensitive to the positive events ratio. Indeed, the mathematical independence between the TSS and the class balance only holds for models that perform equally in every possible case of negative and positive events. For flare forecasting models, the weak performance on samples exhibiting changes in activity results in a direct nonlinear dependency of the TSS on the positive event ratio, and thereby, the class balance.

Let us consider the limit case of a model that perfectly identifies time windows without AC but consistently fails when ACs occur. Such a model can be known as a persistence model. On a given time period, if evaluated on every time window, the persistence model's number of FPs will be equal to the number of FNs, which will be equal to half the number of AC $\frac{\mathrm{C}}{2}$. The number of TPs will be $\mathrm{P} - \frac{\mathrm{C}}{2}$, and the number of TNs will be $\mathrm{N} - \frac{\mathrm{C}}{2}$, where P and N are, respectively, the number of positive and negative events. If we denote $\phi = \frac{\mathrm{P}}{\mathrm{P}+\mathrm{N}}$ the positive event ratio, and $\chi = \frac{\mathrm{C}}{\mathrm{P}+\mathrm{N}}$ the AC rate, the TSS, HSS, and the MCC from Equations (A7), (A10), and (A11) simplify to

$$\mathrm{TSS}_{\mathrm{persistence}} = \mathrm{HSS}_{\mathrm{persistence}} = \mathrm{MCC}_{\mathrm{persistence}} = 1 - \frac{1}{2}\frac{\chi}{\phi(1-\phi)}. \quad \text{(A12)}$$

We may also note that C is at the maximum, equal to twice the minimum between P and N. Therefore, Equation (A12) is defined only when

$$\chi \leqslant \begin{cases} 2\phi & \text{if } \phi \leqslant 0.5 \\ 2(1-\phi) & \text{if } \phi \geqslant 0.5 \end{cases} \text{with } \phi \in [0, 1]. \quad \text{(A13)}$$

Figure 12 shows the plot of a persistence model's TSS, HSS, and MCC, according to Equations (A12) and (A13). The metric linearly increases with the decrease of the AC rate χ, which is typically low in flare forecasting. The performance is nonlinearly dependent on the positive event ratio ϕ, with a stronger sensitivity in most imbalanced cases. For a variation of ϕ from 6% to 12%, the TSS increases from -0.06 to 0.44 with a standard AC rate of 0.12. The model is therefore deemed unskilled in the first case, whereas it can be estimated as mildly proficient in the second case, only because of a doubling of the positive event ratio.

In practice, flare forecasting models can be expected to have a similar TSS, HSS, and MCC sensitivity to ϕ and χ as they have good performance on time windows with the same activity as the previous one, whereas they struggle on time windows with changing activity with respect to the previous one.

The AC rate χ and the positive event ratio ϕ thus emerge as fundamental data set biases in flare forecasting. Our experiment in Appendix B shows that empirically, the sensitivity of the models' metrics to these biases is actually stronger than that of the persistence models. In particular, the TSS seems to exhibit heightened sensitivity to data set biases compared to the HSS, and the MCC appears as the most stable of the three. This is likely due to the positive influence of overcasting on TSS, as well as additional model flaws, such as the low accuracy on C-class negative events in M+ forecasts. These factors further complicate the sensitivity of the models' performance to various data set biases.

A.2.2. TSS's Incomplete Information for Operational Model Purposes

The TSS is inadequately informative in highly imbalanced scenarios, where a good TSS might obscure strong over- or undercasting tendencies, as highlighted by K. D. Leka et al. (2019). While a high TSS ensures accurate classification for most positive and negative events, it does not always translate into a practical model intended for an alarm system if the evaluation is made on an imbalanced set. In cases of substantial imbalance, a high TSS can be reached with a precision close to 0, resulting in a FAR close to 1, rendering the model unfit for an alert system. To illustrate, let us compare the two following models using the same synthetic data set with a positive event ratio of 0.001.

Model 1: $\mathrm{TP} = 99$, $\mathrm{FN} = 1$, $\mathrm{FP} = 5000$, and $\mathrm{TN} = 94{,}900$. Then, $\mathrm{TSS} = 0.94$, $\mathrm{recall} = 0.99$, $\mathrm{precision} = 0.02$, and $\mathrm{F1} = 0.04$. A case similar to this one might arise with X+ flare binary classifiers. A comparative example can be found with the X+ flares binary classifier of X. Huang et al. (2018), achieving a TSS of 0.714 for a FAR of 0.98.

Model 2. $\mathrm{TP} = 90$, $\mathrm{FN} = 10$, $\mathrm{FP} = 50$ and $\mathrm{TN} = 94\ 851$. Then, $\mathrm{TSS} = 0.90$, $\mathrm{recall} = 0.90$ $\mathrm{precision} = 0.64$, and $\mathrm{F1} = 0.75$. This model, while having a lower TSS, maintains a reasonably good recall with significantly higher precision, making it arguably preferable over Model 1 for operational purposes. The TSS contains no information about a model's precision and is, therefore, not a well-suited indicator to select a model for an alarm system in imbalanced cases. Without preferences defined between recall and precision, the F1 score proves more informative in discriminating between a useful and an impractical model in operation. Specific recall and precision preferences can also be considered using the $F\beta$ score, which extends the F1 score by giving β times more importance to the recall than the precision.

A.2.3. HSS Interpretations

The HSS is a weighted harmonic average between the TSS and the MK, with model-dependent importance given to each. The HSS thus synthesizes information about both a model's accuracy and its precision in the different classes, making it arguably a more suitable metric for assessing a model's

suitability as an alarm system than the TSS. However, the model-dependent weight importance between the markedness and the TSS makes it complex to interpret and compare models. The harmonic mean mathematically gives more importance to smaller values. Consequently, for a model tending to undercast the negative class, i.e., negative frequency bias (NFB) smaller than 1, the contribution of the markedness to the HSS is increased. For a model tending to overcast the negative class, i.e., NFB larger than 1, the importance of the TSS contribution to the HSS is conversely increased.

A.2.4. MCC Advantages Over Other Metrics

While the MCC is still uncommon in space weather, it is argued, for general cases, to be a more informative and reliable metric compared to the accuracy and the F1 score (D. Chicco & G. Jurman 2020), the TSS (D. Chicco et al. 2021a), and the HSS (R. Delgado & X.-A. Tibau 2019; D. Chicco et al. 2021b). The MCC is also to be favored over other metrics, such as the area under the curve (AUC) of the receiver operating characteristic (ROC; D. Chicco & G. Jurman 2023) and the Brier score (D. Chicco et al. 2021b), two other metrics of interest in flare forecasts. Empirically, we showed (see Appendix B) the MCC scores to be more resilient to data set composition changes compared to the HSS, which, in turn, is more stable than the TSS. Consequently, the MCC might be preferable for both model selection and comparison across different data sets. It is often a better choice compared to the HSS, as it shares similar information but demonstrates higher stability in extreme cases and is a consistent synthesis of models' class accuracies and precisions. The MCC also allows measuring models' explanatory power agnostically of users' preferences. Despite the MCC's comprehensive assessment of a model's overall quality, the F1 score remains relevant due to its straightforward interpretability for operational alarm system applications. The choice of one metric among the others should ultimately be decided by the importance given to each class and their accuracy and precision. D. Chicco et al. (2021a) surmise that F1 might still be preferred over the MCC when the accurate and confident classification of positive elements holds greater importance than for negative ones. The TSS, on the other end, is still relevant to the balanced problem, or to the imbalanced problem if no importance is given to the model's precision.

Appendix B
Empirical Variability of Standard Metrics to the Evaluation Set Composition

B.1. Standard Metrics

Metrics can be linked with the positive event ratio and the AC rate in a nonlinear way. For instance, the recall and the F1 score of a persistence model are proportional to their ratio: $\mathrm{F1}_{\mathrm{persistence}} = 1 - \frac{\chi}{2\phi}$. In Appendix A.2.1 and Figure 12, we exposed the more complex relationship of the TSS and HSS of a persistence model with these ratios. Similar bias sensitivity should be expected for every model with significant skills deficiency on ACs. To empirically observe the impact of the two ratios on the metrics evaluated on our models, we display the model's performance variation for different combinations of those ratios in Figure 13. The sub-test samples with varying compositions are obtained by varying the start of the test set from 2020 January 1 to 2023 January 1, while maintaining 2023 April 18 as the end date.

All the metrics appear strongly sensitive to the data set composition. The TSS appears to be the most affected, especially in the imbalanced case of the M+ models, while the F1 score and the MCC are the most stable ones. It is worth noting that the F1 score is defined in the range of [0, 1], while the MCC is defined in the range of [−1, 1]. This implies that a unit change in the F1 score corresponds to a double change in magnitude compared to the MCC relative to their respective definition intervals.

B.2. PRSSs

Using the same data set composition variations as presented in the previous section (Appendix B), the variabilities of the PRSSs are displayed in Figure 14.

The PRSSs appear to vary less than their standard metric equivalent. The most resilient ones appear to be the PR-F1 followed by the PR-MCC. The PR-F1 in particular appears remarkably stable except with the C+ model in the most extreme class imbalances, where it becomes slightly positive.

With the exception of the PR-TSS, the PRSSs indicate a consistent lack of performance in the M+ case and null to slightly positive for the C+ model. This suggests that despite the strong impact of the data set biases on the performance evaluation, models reaching state-of-the-art performance could be expected to consistently struggle to outperform persistence models over varying subsets of the solar cycle.

ORCID iDs

G. Francisco https://orcid.org/0000-0003-3694-7813
M. Berretti https://orcid.org/0009-0007-2465-1931
S. Chierichini https://orcid.org/0009-0005-6746-2917
R. Mugatwala https://orcid.org/0000-0003-4443-9966
J. Fernandes https://orcid.org/0000-0002-1663-3334
T. Barata https://orcid.org/0000-0001-6106-8285
D. Del Moro https://orcid.org/0000-0003-2500-5054

References

Abadi, M., Agarwal, A., Barham, P., et al. 2015, TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems, https://www.tensorflow.org/
Aghajanyan, A., Zettlemoyer, L., & Gupta, S. 2020, arXiv:2012.13255
Athiray, P. S., & Winebarger, A. R. 2024, ApJ, 961, 181
Barnes, G., Leka, K. D., Schrijver, C. J., et al. 2016, ApJ, 829, 89
Barnes, W. T., Cheung, M. C. M., Bobra, M. G., et al. 2020, JOSS, 5, 2801
Bloomfield, D. S., Higgins, P. A., McAteer, R. T. J., & Gallagher, P. T. 2012, ApJL, 747, L41
Brown, E. J. E., Svoboda, F., Meredith, N. P., Lane, N., & Horne, R. B. 2022, SpWea, 20, 3
Camporeale, E. 2019, SpWea, 17, 1166
Chicco, D., & Jurman, G. 2020, BMCG, 21, 6
Chicco, D., & Jurman, G. 2023, Biomed. Data Min., 16, 4
Chicco, D., Tötsch, N., & Jurman, G. 2021a, Biomed. Data Min., 14, 13
Chicco, D., Warrens, M., & Jurman, G. 2021b, IEEA, 9, 78368
Cinto, T., Gradvohl, A. L. S., Coelho, G. P., & da Silva, A. E. A. 2020, MNRAS, 495, 3332
Crocker, J. C., & Grier, D. G. 1996, JCIS, 179, 298
Delgado, R., & Tibau, X.-A. 2019, PLoSO, 14, e0222916
Deng, J., Dong, W., Socher, R., et al. 2009, in 2009 IEEE Conf. on Computer Vision and Pattern Recognition (Piscataway, NJ: IEEE), 248
Deng, Z., Wang, F., Deng, H., et al. 2021, ApJ, 922, 232
Deshmukh, V., Flyer, N., van der Sande, K., & Berger, T. 2022, ApJS, 260, 9
Francisco, G. 2025, Patch-CNN for Weakly-supervised Flare Forecasting, v1, Zenodo, doi:10.5281/zenodo.14790146

Francisco, G., Del Moro, D., Barata, T., & Fernandes, J. 2024, SDO 2H Machine Learning Dataset v2, Zenodo, doi:10.5281/zenodo.10465436
Garcia, H. A. 1994, SoPh, 154, 275
Guastavino, S., Marchetti, F., Benvenuto, F., Campi, C., & Piana, M. 2022a, A&A, 662, A105
Guastavino, S., Marchetti, F., Benvenuto, F., Campi, C., & Piana, M. 2022b, FrASS, 9, 399
Guastavino, S., Piana, M., & Benvenuto, F. 2022, in IEEE Transactions on Neural Networks and Learning Systems, 35 (Piscataway, NJ: IEEE), 1993
Hanssen, A., & Kuipers, W. 1965, On the Relationship between the Frequency of Rain and Various Meteorological Parameters 58 (Koninklijk Nederlands Meteorologisch Instituut)
Heidke, P. 1926, GeAn, 8, 301
Hu, E. J., Shen, Y., Wallis, P., et al. 2021, arXiv:2106.09685
Huang, X., Wang, H., Xu, L., et al. 2018, ApJ, 856, 7
Kingma, D. P., & Ba, J. 2014, arXiv:1412.6980
Kokkinos, I. 2017, in 2017 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) (Piscataway, NJ: IEEE), 5454
Leka, K. D., Park, S.-H., Kusano, K., et al. 2019, ApJS, 243, 36
Lemen, J. R., Title, A. M., Akin, D. J., et al. 2012, SoPh, 275, 17
Li, X., Zheng, Y., Wang, X., & Wang, L. 2020, ApJ, 891, 10
Loshchilov, I., & Hutter, F. 2017, arXiv:1711.05101
Matthews, B. 1975, BBAcB, 405, 442
Nishizuka, N., Sugiura, K., Kubo, Y., Den, M., & Ishii, M. 2018, ApJ, 858, 113
Pandey, C., Angryk, R. A., & Aydin, B. 2023, in Machine Learning and Knowledge Discovery in Databases: Applied Data Science and Demo Track. ECML PKDD 2023, Vol. 14175 ed. G. De Francisci Morales (Cham: Springer), 72
Park, E., Moon, Y.-J., Shin, S., et al. 2018, ApJ, 869, 91
Peirce, C. S. 1884, Sci, 4, 453
Pesnell, W. D., Thompson, B. J., & Chamberlin, P. C. 2012, SoPh, 275, 3
Plutino, N., Berrilli, F., Del Moro, D., & Giovannelli, L. 2023, AdSpR, 71, 2048
Plutino, N., Michele, B., Grégoire, F., et al. 2024, Solar Flare Catalog: Plutino Extension, v2, Zenodo, doi:10.5281/zenodo.11150339
Ruder, S. 2017, arXiv:1706.05098
Schrijver, C. J. 2009, AdSpR, 43, 739
Selvaraju, R. R., Cogswell, M., Das, A., et al. 2016, arXiv:1610.02391
Shazeer, N., Mirhoseini, A., Maziarz, K., et al. 2017, arXiv:1701.06538
Springenberg, J. T., Dosovitskiy, A., Brox, T., & Riedmiller, M. 2014, arXiv:1412.6806
Sun, Z., Bobra2, M. G., Wang, X., et al. 2022, ApJ, 931, 23
Tan, M., & Le, Q. V. 2021, arXiv:2104.00298
Van der Sande, K., Flyer, N., Berger, T. E., & Gagnon, R. 2022, FrASS, 9, 354
Yi, K., Moon, Y.-J., Lim, D., Park, E., & Lee, H. 2021, ApJ, 910, 8