The Astrophysical Journal, 999:87 (16pp), 2026 March 1 https://doi.org/10.3847/1538-4357/ae3f26

Investigating the Efficacy of Topologically Derived Time Series for Flare Forecasting. II. XGBoost Model

Thomas Williams[1], Christopher B. Prior[1], David MacTaggart[2], O. P. M. Aslam[2], and D. Shaun Bloomfield[3]
[1] Department of Mathematical Sciences, Durham University, Durham, UK; tomwilliamsphd@gmail.com
[2] School of Mathematics & Statistics, University of Glasgow, Glasgow, UK
[3] School of Engineering, Physics and Mathematics, Northumbria University, Newcastle upon Tyne, UK
Received 2025 December 16; revised 2026 January 23; accepted 2026 January 27; published 2026 February 25

Abstract

Solar flares are a primary driver of space weather, and forecasting their occurrence remains a significant challenge. This paper presents a novel flare prediction model based on topologically derived photospheric magnetic parameters. We employ the ARTop framework to compute the time-dependent input rates of magnetic winding and helicity across $>10^5$ active region observations, decomposing them into current-carrying and potential components to reduce sensitivity to optical flow methods. An XGBoost machine learning model is trained on these time series, alongside engineered features including rolling statistics, kurtosis, and flare history, to predict the probability of $\geqslant$M1.0 class flares within the next 24 hr. The model demonstrates strong performance on a validation set, with a true skill statistic (TSS) of 0.804 for once-daily operational region forecasts. When applied to a fully independent dataset, the model achieves a TSS of 0.524. A SHAP analysis confirms the model's physical interpretability, identifying flare history and accumulated current-carrying winding and helicity as the most important features. The main challenges identified are false positives arising from active regions with frequent C-class flaring and systematic errors via projection effects when active regions are near the limb. Excluding limb-affected data yields no improvement in the holdout set TSS (0.521 versus 0.524), due to the overall decreased number of flares. However, our per-region analysis indicates that mitigating these projection effects is crucial for future operational deployment. This work establishes magnetic topology, particularly its current-carrying components, as a highly effective and physically meaningful set of predictors for solar flare forecasting.

Unified Astronomy Thesaurus concepts: Solar flares (1496); Space weather (2037); Solar active region magnetic fields (1975); Astronomy databases (83); Classification (1907); Neural networks (1933)

Materials only available in the online version of record: machine-readable table

1. Introduction

Solar flares are intense, impulsive releases of magnetic energy in the solar corona, characterized by sudden enhancements in electromagnetic radiation across the spectrum—from radio waves through ultraviolet (UV) to X-rays and gamma rays. These events originate in magnetically complex active regions and produce copious nonthermal electron and ion acceleration, which drive phenomena such as solar radio bursts and EUV/X-ray enhancements (A. O. Benz 2017). The rapid ionization of Earth's lower ionosphere (notably the D and E layers) produces sudden ionospheric disturbances, often referred to as the magnetic crochet effect, which can severely degrade high-frequency radio communication and disrupt Global Navigation Satellite System signals (M. Temmer 2021; N. Gopalswamy 2022). In addition, flare-associated enhancements in ultraviolet and soft X-ray flux increase atmospheric heating, expanding the upper atmosphere and enhancing drag on low-Earth-orbit satellites, thereby affecting satellite orbit maintenance and lifetime (N. Gopalswamy 2022).

In the broader landscape of space weather, solar flares frequently occur with coronal mass ejections (CMEs) and solar energetic particle (SEP) events to produce compounded effects in near-Earth space. While CMEs drive geomagnetic storms via magnetospheric interactions, flares contribute prompt injections of high-energy particles that may arrive within tens of minutes, posing acute risks to satellite electronics, astronaut radiation exposure, and commercial aviation at high latitudes. Notably, SEP events accelerate via both flare-associated reconnection and CME-driven shocks, underscoring the necessity to model both mechanisms in operational forecasting (K.-L. Klein & S. Dalla 2017). Subsequently, solar flares constitute critical drivers of electromagnetic, ionospheric, and radiative disruptions in the Sun–Earth system and remain primary targets for predictive machine learning (ML) pipelines in operational space weather models.

The flux of magnetic helicity has long been considered an important quantity in the study of active regions and solar flares (A. A. Pevtsov et al. 2003; S.-H. Park et al. 2008; P. Vemareddy 2021; M. Korsós et al. 2022; Y. Liu et al. 2023). Another similar quantity, known as the magnetic winding, which describes the geometric twisting of more horizontal magnetic field lines, is a relativity novel quantity that has also been shown to have additional predictive efficacy (C. Prior & D. MacTaggart 2020; D. MacTaggart et al. 2021; B. Raphaldini et al. 2022; O. Aslam et al. 2024).

T. Williams et al. (2025, hereafter WPM) leverage the Active Region Topology (ARTop) framework to generate time-resolved metrics of magnetic helicity flux $\mathcal{H}'$ and winding input $\mathcal{L}'$ rates derived from Helioseismic and Magnetic Imager (HMI; J. Schou et al. 2012) data aboard the Solar Dynamics Observatory (SDO) in the form of Space-Weather HMI Active Region Patches (SHARP). The aim of this study is to assess the predictive potential of topological quantities with respect to solar flares. WPM focuses on imbalance measures—specifically the absolute difference in current-carrying versus potential components of topological inputs ($\delta\mathcal{L}'$ and $\delta\mathcal{H}'$)—which were shown to be comparatively insensitive to velocity-smoothing choices and correlate strongly with flaring activity. Through analysis of a comprehensive dataset of 144 active regions (including both flaring and nonflaring cases), the study revealed that spikes in $\delta\mathcal{L}'$ and $\delta\mathcal{H}'$ time series—including combinations with line-of-sight velocity ($v_z\ \delta\mathcal{L}_c{}'$ and $v_z\ \delta\mathcal{H}_c{}'$)—often precede M- and X-class flares by up to several hours, with statistically significant deviations beyond 2σ envelopes in their running mean. These findings suggested that topologically derived signals, especially those representing current-carrying flux imbalance, offer promising early indicators of major solar eruptive events and warrant further investigation in solar flare forecasting models.

In recent years, numerous studies have adopted an ML approach in attempts to predict solar flares (N. Nishizuka et al. 2017, 2021; S. Sinha et al. 2022; K. Yi & Y.-J. Moon 2022; J. Yun & J. Shin 2024; M. Li et al. 2025). For example, M. G. Bobra & S. Couvidat (2015) utilize 4 yr' of HMI SHARP data to train a support vector machine (SVM) classifier that is trained on 25 derived features of active regions from which they conclude only a few of these parameters are needed to predict flaring. Similarly, K. Florios et al. (2018) compare SVM, random forest (RF), and multilayer perceptron methods across 5 yr of Solar Cycle 24. Their findings are that RF performs best, and when flares exceeding M1.0 are considered, they obtain a true skill statistic (TSS) score of 0.74 provided a probability threshold of 0.15 is specified for a dataset akin to a holdout set. While other methods adopt convolutional neural networks (CNNs; K. van der Sande & A. Muñoz-Jaramillo 2025), recurrent neural networks (C. Pandey et al. 2023) or some combination of the two (S. Guastavino et al. 2022) for the basis of their predictive models, these studies incorporate advanced ML frameworks explicitly targeting forecasting capability instead of mere pattern recognition and align their findings with operational space weather goals such as 6–72 hr flare prediction and full-disk or near-real-time applicability on real solar data. Though as is common in all flare prediction studies, they suffer from highly imbalanced data that can impact model training and skew skill scores and model evaluation.

Another algorithm rooted in advanced ML frameworks in the Extreme Gradient Boosting (XGBoost) method is a high-performance, scalable implementation of gradient boosting decision trees that has gained widespread adoption across various ML tasks, including time series forecasting. It builds an ensemble of decision trees in a sequential manner that are optimized and support both first- and second-order gradient information, including regularization to control model complexity (T. Chen & C. Guestrin 2016). For time series and classification tasks—such as forecasting solar flares based on SHARP parameters—XGBoost has demonstrated robust performance by effectively handling nonlinear relationships, missing data, and feature importance interpretation (T. Cinto et al. 2020). Similarly, a recent comparative study by J. Bringewald & O. Parisot (2025) found that XGBoost performed strongly compared to RF and k-nearest neighbor models when classifying solar flares into GOES categories (B, C, M, X), and that XGBoost benefits significantly from increased dimensionality with "state-of-the-art" performance across numerous data science tasks (T. Chen & C. Guestrin 2016).

For the purpose of forecasting solar flares, we will define explicitly what we constitute as a prediction. Historically, predictive models have utilized a broad range of time frames between 6 and 48 hr (see, for example, Table 2 in M. G. Bobra & S. Couvidat 2015). However, operationally, these forecasts or predictions on the likelihood of flaring are made once per day and give the outcome for the following 24 hr. As such, more recent predictive methods now align with these constraints (S. Sinha et al. 2022; V. Deshmukh et al. 2023; C. Pandey et al. 2023, for example). It is with these operational capabilities in mind that we will also adopt a 24 hr cadence (at UTC midnight) for making predictions of flaring for the upcoming day.

This second paper focuses on the creation of a supervised ML model with XGBoost to predict the likelihood of M- and X-class flaring within the next 24 hr using the dataset outlined in WPM. In Section 2, the ARTop code and living dataset, and data preparation employed for this study are outlined. Section 3 discusses the XGBoost model, the feature engineering, hyperparameter and model tuning, and the model evaluation metrics, while the model is evaluated on validation data in Section 4. Our main results and analysis are presented in Section 5 and concluding remarks are provided in Section 6.

2. Dataset

2.1. ARTop

ARTop is an open-source tool for studying the input of topological quantities into solar active regions at the photospheric level. ARTop utilizes vector magnetograms (J. T. Hoeksema et al. 2014) from HMI SHARP data to create maps, time series, and other metrics derived from input rates of magnetic helicity and magnetic winding fluxes (see C. Prior & D. MacTaggart 2020 for a detailed summary of their meaning and importance in solar applications). Figure 1 illustrates the geometrical and physical interpretation of these fluxes. Motions of field lines at the solar surface are estimated from the magnetogram data and the Differential Affine Velocity Estimator for Vector Magnetograms (DAVE4VM) code (P. W. Schuck 2008) for the plasma velocity. This is used to estimate the mutual rotation of individual and mutual pairs of field lines at the solar surface, due to either in-plane plasma motion braiding the field, or the apparent field line motion due to field lines with complex topology emerging from the solar interior/submerging from the corona. The winding flux distribution $\mathcal{L}'$ at a given point in the photosphere is the average rotational motion of *all* field lines around that point. The helicity flux is the winding flux weighted by the field's B_z flux (field strength weighted winding). The quantities $\delta\mathcal{L}'$ and $\delta\mathcal{H}'$ are calculated from these distributions as described above. The crucial difference between the winding and helicity-related quantities is that the winding quantities tend to be dominated by the field around the polarity inversion line(s),

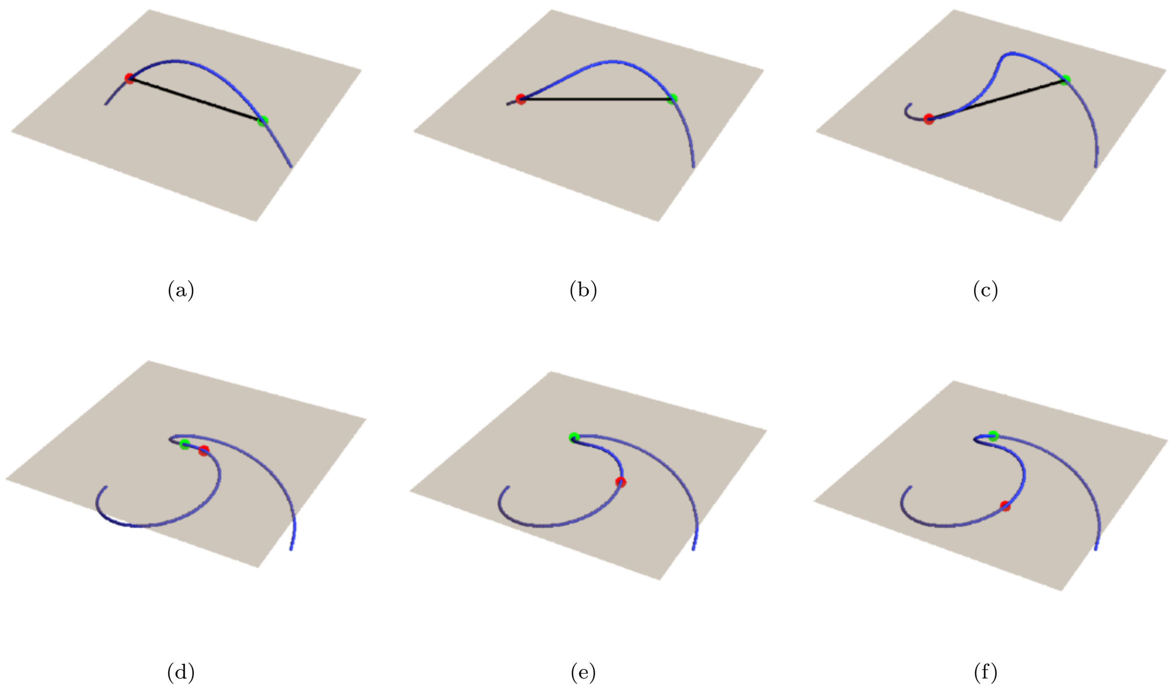

Figure 1. Illustrations of field properties measured by the winding flux $\mathcal{L}'$ and helicity flux $\mathcal{H}'$. Here, the photosphere is represented as a 2D plane (gray) with the protruding magnetic field line shown in blue. The red and green spheres indicate the points of the magnetic field at the photospheric level; the rotation of the black line joining them in the plane indicates their relative rotation, which is measured by the winding and helicity fluxes. Panels (a)–(c) depict the entanglement of a field line due to a rotational fluid motion at the photosphere (depicted as a plane). The winding measures the rotation of the green and red points where the field line pierces the plane. Panels (d)–(f) depict the rotational motion evaluated by the winding due to a contorted field line emerging through the photosphere. The helicity measures these rotations multiplied by the flux in the field lines.

while the helicity is dominated more by the field at the major footpoints (sunspots). For more comprehensive definitions of these quantities and the practical matter of their calculation in ARTop, please refer to WPM and K. Alielden et al. (2023), and references therein.

One of the main results from WPM is that the decomposition of magnetic winding (and magnetic helicity) into components for current-carrying and potential topology is sensitive to the velocity smoothing employed in the DAVE4VM code. Given the lack of sensitivity of $\delta\mathcal{L}'$ to these parameters, we instead split $\delta\mathcal{L}'$ and $\delta\mathcal{H}'$ into positive and negative components, whereby positive (negative) would indicate that the dominant input of topology at the photospheric level is current-carrying (potential) topology. This is achieved by the following conditional expressions:

$$\delta\mathcal{L}_c' = \begin{cases} \delta\mathcal{L}', & \text{if } \delta\mathcal{L}' > 0 \\ 0, & \text{otherwise,} \end{cases}$$
$$\delta\mathcal{H}_c' = \begin{cases} \delta\mathcal{H}', & \text{if } \delta\mathcal{H}' > 0 \\ 0, & \text{otherwise,} \end{cases} \quad (1)$$

and

$$\delta\mathcal{L}_p' = \begin{cases} \delta\mathcal{L}', & \text{if } \delta\mathcal{L}' < 0 \\ 0, & \text{otherwise,} \end{cases}$$
$$\delta\mathcal{H}_p' = \begin{cases} \delta\mathcal{H}', & \text{if } \delta\mathcal{H}' < 0 \\ 0, & \text{otherwise.} \end{cases} \quad (2)$$

This subsequently reduces the sensitivity of the velocity-smoothing profile selected, while still allowing for current-carrying decompositions to be investigated. These current-carrying components of the helicity and winding may then be incorporated alongside the other topological quantities of our dataset.

2.2. Living Dataset

The 232 SHARP regions utilized in this study are curated from the dataset outlined in WPM and A. Hollanda et al. (2021). We ensured a number of X-class flaring regions are included in this curation as they are relatively rare among the dataset as a whole, but form an important target for predictive purposes. Other regions included in this study have been sampled at intervals such as to allow an even coverage across the solar cycle. As the dataset developed for this manuscript is a living dataset, it will continuously evolve as more of the regions outlined by A. Hollanda et al. (2021) are processed with ARTop. The reason 232 SHARP regions are selected for this study is the result of the computing resources required during curation with ARTop. As is outlined in WPM, each active region is sampled with a cadence of 720 s for the entire duration it is visible within the HMI data. This cadence is needed as ARTop requires the full time series in order to obtain accurate integrated quantities. Additionally, ARTop must create 2D vector maps for velocity before it may then determine the topological quantities for each pixel of the SHARP. These calculations are expensive, with the computational costs growing exponentially for larger regions. For example, large regions like SHARP 5298 that travel across the width of the solar disk can take over a week to process on a dedicated 32-core machine. As such, time constraints dictate our sample size for this study, as was the case in WPM, where the total number of SHARP regions was 144.

Table 1
SHARP Regions That Form the Validation Set for the `XGBoost` Model

NOAA Active Region	SHARP Number	Strongest Flare	# X-class Flares	# M-class Flares	# C-class Flares
11070	14	…	…	…	…
11132	285	…	…	…	…
11158	377	X2.2	1	5	54
11211	590	…	…	…	…
11214/17	602	…	…	…	…
11291	851	…	…	…	…
11429/30	1449	X5.4	3	14	36
11437	1480	…	…	…	…
11469/73	1611	C2.6	…	…	12
11488	1688	…	…	…	…
11546	1942	…	…	…	…
11561	1997	…	…	…	…
11572	2036	…	…	…	…
11640	2337	C4.0	…	…	6
11829	3113	C1.6	…	…	1
11839	3171	…	…	…	…
11893/900	3364	X1.0	1	4	33
11910	3441	…	…	…	…
11922	3490	…	…	…	…
12003/06	3845	C7.3	…	…	10
12111	4328	…	…	…	…
12297	5298	X2.1	1	22	92
12661	7034	C8.0	…	…	7
12673	7115	X9.3	4	26	52
12683	7148	…	…	…	…

Of the 232 SHARP regions analyzed here, 110 regions are considered to be flaring (C-class and above), with a further 122 nonflaring regions (no flare, A- and B-classes). In all, the dataset presented in this manuscript contains 2142 X-ray flares of magnitude C1.0 and above. However, in this work, a region is only considered to be flaring if flares of X-ray class magnitude M1.0 and above are recorded in the Heliophysics Event Knowledgebase (HEK; N. E. Hurlburt 2022). As such, this means we have 57 and 175 flaring and nonflaring active regions with a total of 384 M- and X-class flares. A list of the SHARP regions that form the validation and holdout sets is presented in Tables 1 and 2. These contain 25 and 12 regions with their respective strongest flare magnitudes and number of C-, M-, and X-class flares. The same information is provided in the Appendix for the training set, which is formed of 207 regions.

2.3. Data Preparation

The data used in this study are derived from the living dataset outlined in Section 2.2. All SHARP active regions are loaded into memory, cleaned, and separated into training and validation subsets. Importantly, a stratified sampling scheme is applied so that the validation set contains the three benchmark major flaring regions (NOAA active regions 11158, 11429, and 12673) from WPM. This ensures that the validation set is both statistically independent and astrophysically representative of high-impact events. The full list can be seen in Table 1.

Table 2
SHARP Regions That Form the Holdout Set for the `XGBoost` Model

NOAA Active Region	SHARP Number	Strongest Flare	# X-class Flares	# M-class Flares	# C-class Flares
11976/80	3730	M3.0	…	1	9
12047/51	4071	M1.8	…	2	23
12228	4896	…	…	…	…
12282	5186	M2.4	…	1	6
12322	5446	M4.0	…	7	6
12449/50	6078	M3.9	…	1	6
12740	7357	C9.9	…	…	9
13473	10307	C2.4	…	…	1
13841	11968	M1.5	…	4	8
13920	12392	M2.7	…	1	21
13922	12404	M2.2	…	8	15
13924	12405	M2.1	…	2	11

In an attempt to reduce the impact of spurious measurements associated with projection effects at the limb, all snapshots where the heliographic position of the SHARP region exceeds a defined longitudinal threshold ($\pm 60°$) are flagged based on the percentage of pixels affected so as to conform with the findings of M. G. Bobra et al. (2014). The idea being that this may allow our ML model, which is created in Python, to identify trends in the spurious values in the `ARTop` calculations during training, which we will discuss later in Section 5. Additionally, rows containing missing or corrupted values are removed, and the remaining data for each region are concatenated into unified training and validation `DataFrames`, which are subsequently stored in serialized pickle format for reproducibility and efficient downstream use. At this stage, the percentage of NaN entries and limb-contaminated samples is reported to quantify data quality (typically <6% and <25% of total samples, respectively).

3. `XGBoost` Model

This section outlines the supervised ML pipeline designed to forecast solar flares within a 24 hr prediction window. The pipeline integrates a series of custom data preparation routines, feature engineering procedures, and classification using the `XGBoost` algorithm (T. Chen & C. Guestrin 2016). The modular framework is composed of (1) data extraction and organization, (2) feature engineering from SHARP vector magnetogram time series, (3) dataset partitioning into training and validation subsets, (4) model training with hyperparameter optimization, and (5) evaluation using multiple performance metrics with optimized classification thresholds. The overall workflow of the model is illustrated in Figure 2.

3.1. Feature Engineering

A central component of the pipeline is the construction of features designed to capture both the instantaneous state and temporal evolution of the active region magnetic field by creating rolling statistics, lagged variables, and higher-order temporal descriptors. For each SHARP-derived quantity, rolling means and standard deviations are computed, capturing

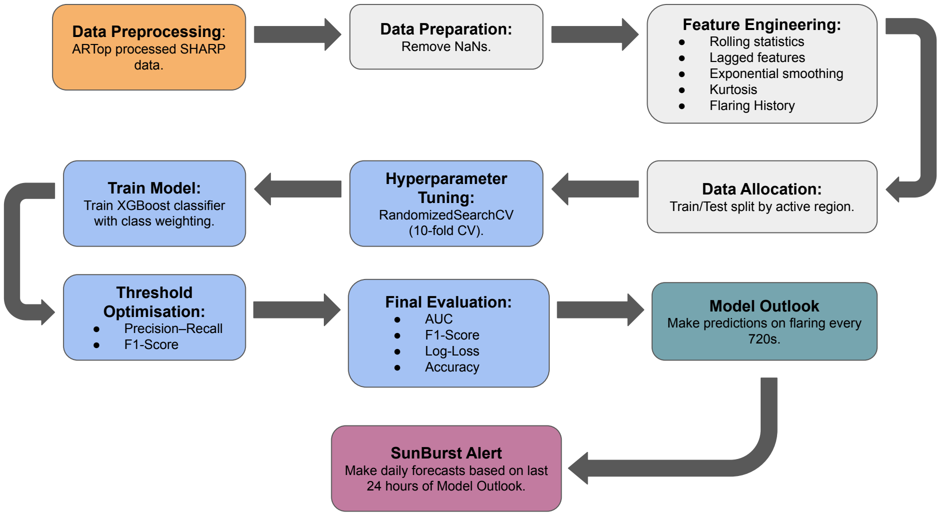

Figure 2. Workflow diagram of the `XGBoost` classification model pipeline.

short-term variability. These parameters provide the model with memory of past activity.

As outlined by WPM, the quantities given by condition (1) (particularly $\delta\mathcal{L}_c{}'$) provide promising correlations with flares from 65 flaring active regions. As the living dataset does not contain these quantities, they are created in place of the other winding and helicity rates. Similarly, in WPM, a buildup of accumulated winding and helicity seems to indicate a greater propensity for larger flares to occur within an active region. As these accumulations are based upon $\delta\mathcal{L}'$ and $\delta\mathcal{H}'$, when potential-field topology is dominant for periods, this can reduce the accumulated quantities and make it more difficult to estimate the buildup of complex current-carrying topology. For this reason, the accumulated helicity and winding are also converted into the current-carrying components only. As such, the quantities of the living dataset that are incorporated into our dataset for ML are: $\delta\mathcal{L}_c{}'$ and $\delta\mathcal{H}_c{}'$ (and their accumulated values), $\delta\mathcal{L}'$, $\delta\mathcal{H}'$, $v_z\,\delta\mathcal{L}_c{}'$, $v_z\,\delta\mathcal{H}_c{}'$, and $v_z|B_z|$. Additional lagged data indicating the previous behavior at offsets of 0.2 and 0.4 hr alongside rolling statistics for the mean and standard deviation taken over the last 2 hr for each quantity are also computed.

To further smooth the high-cadence time series, single exponential smoothing (SES; R. G. Brown 1959; P. R. Winters 1960; C. C. Holt 2004) is applied. These operations suppress stochastic noise while preserving physically meaningful trends. The advantage of SES compared to a standard moving mean when removing high-frequency noise is that all previous points are not weighted equally but by utilizing exponentially decaying weighting factors such that more recent time series values are weighted more heavily. This weighting allows for more recent values to be deemed more important than those seen at the start of the smoothing window, giving a recency bias. This choice is motivated by the fact that active region properties evolve on timescales comparable to the 12–24 hr forecast windows considered, while several quantities used in this study exhibit increased variability in the period leading up to flare onset (WPM). SES reduces the influence of older observations while retaining short-term temporal structure, allowing the model to emphasize recent evolution without explicitly differentiating noisy time series. This is particularly appropriate given the cadence and measurement uncertainties associated with vector magnetogram-derived parameters.

In addition to smoothing, higher-order statistics are extracted. For example, rapid increases in the excess kurtosis have been shown to be an important precursor to system bifurcations (or critical transitions) in a number of fields such as in ecological systems (V. Dakos et al. 2019), climate (N. Boers 2018), the economy (C. Sevim et al. 2014), and medicine (L. Chen et al. 2012). As such, kurtosis values are determined for key magnetic quantities (e.g., $\delta\mathcal{L}_c{}'$ and $\delta\mathcal{H}_c{}'$) over 3, 6, and 12 hr windows, providing a measure of intermittency and "burstiness" in the magnetic field evolution. These descriptors were shown in WPM to correlate with flare productivity.

Furthermore, we engineer flare history features: backward-looking flare indices are calculated for the 12 and 24 hr preceding a given observation. These features quantify an active region's recent flaring history, which is often predictive of future activity. In order to provide a proxy of this historical energy expenditure of flares within an active region, time series that quantify a score for the flaring are created. The idea being that over a period of several hours, an active region could release similar quantities of built-up magnetic energy by numerous smaller C- and M-class flares as it might with a single X-class flare. To assign a numerical value to a flare ($f_{\rm val}$), we follow a logarithmic classification function such that

$$f_{\rm val} = \begin{cases} 10x, & \text{if C-class} \\ 100x, & \text{if M-class} \\ 1000x, & \text{if X-class} \\ 0, & \text{Otherwise}, \end{cases} \tag{3}$$

where x is the flare classification's numerical suffix. For example, a C2.6 flare would be assigned a value of 26 while an X9.3 flare would be assigned a value of 9300. The motivation for taking a multiplier of 10 on the C-class flares allows for greater separation between the flaring and nonflaring outcomes, which may aid model training. Based upon this, the flare score time series are then calculated on running-sums of all the $f_{\rm val}$ values utilizing kernels of the previous 24 and 12 hr, respectively. This is based on a common flare score metric (A. Antalova 1996; V. Abramenko 2005; S.-h. Park et al. 2010).

Note that here, the ranges do not include the last hour to approximate real-time observational delays between a flare being detected and its magnitude being quantified and publicized.

To stabilize feature distributions, all variables with the exception of integrated quantities, flare labels, limb-affected pixel percentage, and kurtoses are log-transformed. Negative values are preserved by mapping their magnitudes through a signed $\log_{10}$ transform (Equation (4)), ensuring the statistical properties of positive and negative excursions are comparably scaled. We chose this transformation as it reduces skew and prevents dominance of a small subset of parameters with large raw magnitudes. Mathematically, the expression is as follows:

$$X = \begin{cases} -\log_{10}|X|, & \text{if } X < 0 \\ \log_{10} X, & \text{if } X > 0 \\ 0, & \text{otherwise,} \end{cases} \tag{4}$$

where X is the metric to normalize.

Finally, forward-looking flare labels are constructed at 24 hr horizons, defining the binary classification target variable.

3.2. Dataset Partitioning

Following feature engineering, the living dataset is organized into training and validation sets that contain 90% and 10% of the SHARP regions, respectively. It is also worth noting that each set retained the full time series for its assigned SHARP regions to prevent temporal leakage, ensuring that validation and holdout sets purely consist of unseen data. The regions included in the validation set are outlined in Table 1, while the training set is included in the Appendix. An additional holdout set is also included (Table 2) that consists of 12 active regions that are not part of the 232 SHARP regions, which form the training and validation sets. Within each set, flare magnitudes are binarized: samples with no flare or C-class flare activity are labeled as 0 (nonflaring), while those associated with M- or X-class flares within the next 24 hr are labeled as 1 (flaring). This decision reflects operational needs to predict impactful space weather events, where false negatives for strong flares are particularly costly.

3.3. Evaluation Metrics

The final model is assessed on the validation set using five metrics. Letting TN, TP, FN, and FP denote true negative, true positive, false negative, and false positive outcomes, respectively, the metrics are

1. F1 score, defined as the harmonic mean of precision and recall:

$$\mathrm{F1} = 2 \cdot \frac{\mathrm{Precision} \cdot \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}}, \tag{5}$$

with

$$\mathrm{Precision} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}, \quad \mathrm{Recall} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}.$$

2. Area under the ROC curve (AUC) in terms of the true positive rate (TPR) and false positive rate (FPR) is computed as

$$\mathrm{AUC} = \int_0^1 \mathrm{TPR}(\mathrm{FPR})\, d(\mathrm{FPR}), \tag{6}$$

where $\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}$ and $\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{FP}+\mathrm{TN}}$. It is worth noting that TPR and Recall are the same metric but are named differently in the literature when referenced in F1 scores compared to other metrics.

3. Log-loss, a measure of probability calibration:

$$\mathrm{LogLoss} = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(p_i) + (1 - y_i)\log(1 - p_i)], \tag{7}$$

where $y_i \in \{0, 1\}$ are true labels and p_i are predicted probabilities.

4. Accuracy, the overall proportion of correctly classified instances:

$$\mathrm{Accuracy} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}}. \tag{8}$$

5. TSS, also known as the Hansen–Kuiper skill score, which measures the ability to discriminate between events and non-events while accounting for imbalanced datasets:

$$\mathrm{TSS} = \mathrm{TPR} - \mathrm{FPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} - \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}}, \tag{9}$$

where TSS ranges from -1 to $+1$, with $+1$ indicating perfect prediction and 0 indicating no skill.

These metrics are chosen to provide complementary perspectives: the F1 score highlights skill at rare-event detection, AUC measures discriminative ability across thresholds, log-loss evaluates probabilistic calibration, accuracy captures overall classification success, and TSS provides a threshold-independent measure of forecast skill that is particularly valuable for imbalanced datasets common in flare forecasting. Reporting this suite of metrics ensures comparability with other recent flare forecasting studies (e.g., N. Nishizuka et al. 2017; T. Cinto et al. 2020).

3.4. Model Training and Optimization

We utilize the `XGBoost` classifier, a decision-tree ensemble method that constructs boosted trees in sequence, each correcting the errors of the prior ensemble. The algorithm is well suited for structured tabular data and provides state-of-the-art performance in many classification tasks (T. Chen & C. Guestrin 2016). The classifier is configured with a logistic objective function (`binary:logistic`) and AUC and log-loss evaluation metrics. As flaring intervals only constitute a small minority of the dataset (≈8%), we apply a positive class weight factor (`scale_pos_weight`) to account for the data imbalance. This penalizes misclassification of flares relative to nonflaring intervals, mitigating bias toward the majority class, i.e., the nonflaring intervals.

Hyperparameter tuning is conducted using the `RandomizedSearchCV` procedure from `scikit-learn` (F. Pedregosa et al. 2011). Unlike `GridSearchCV`, which exhaustively explores all combinations, randomized search samples from predefined distributions, enabling memory-efficient exploration of high-dimensional parameter spaces. We define distributions for the number of estimators (`n_estimators`: 300–600), tree depth (`max_depth`: 6–10), learning rate (`learning_rate`: uniform in [0.05, 0.25]), subsampling fraction (`subsample`: uniform in [0.7, 1.0]), and minimum child weight (`min_child_weight`: 1–10). A total of 500 configurations are sampled for each fold of the search. From this, a stratified tenfold cross-validation scheme is employed to ensure balanced representation of flaring and nonflaring samples across folds. Model performance is evaluated using three complementary metrics: AUC, log-loss, and F1 score. To select the optimal hyperparameter set, we introduce a composite refit criterion. Each metric is normalized to the range [0, 1], and a weighted average computed by

$$S = 0.4 \cdot \widehat{\mathrm{AUC}} + 0.2 \cdot \widehat{\mathrm{LogLoss}} + 0.4 \cdot \widehat{\mathrm{F1}}, \tag{10}$$

where hats denote normalized scores. This composite score balances ranking ability, probability calibration, and rare-event sensitivity. It is worth noting here that the weighting adopted for metrics that comprise S was determined through trial and error.

As mentioned earlier, the target variable is set using binary logic for whether there is flaring exceeding a GOES magnitude of M1.0 within the next 24 hr. As `XGBoost` produces probabilistic outputs, a decision threshold must be applied to classify events. Rather than adopting the default threshold of 0.5, we optimize the threshold on the validation set by computing precision–recall curves and select the value that maximizes the F1 score. This ensures that the final classification strikes a balance between minimizing false alarms (precision) and maximizing flare detection (recall). The optimal threshold of 0.476 is stored alongside the trained model for operational use, and has a value closer to the default 0.5, such as used in T. Cinto et al. (2020) and N. Nishizuka et al. (2021) (0.5 and 0.4), than is deployed in other models such as K. Florios et al. (2018) (0.15).

3.5. Operational Deployment

For each observation of a SHARP region, the `XGBoost` model is run and makes a prediction on the likelihood of an M1.0+ flare occurring in the 24 hr following that SHARP observation. For the purpose of clarity, we will henceforth refer to these instantaneous predictions as the Model Outlook. The reason for this clarification is that we also make forecasts once per UT day about whether flaring will occur the following UT day. This cadence is selected so as to be in line with how operational forecasts on space weather are made. These daily forecasts are the final outcome from the pipeline discussed here, and are subsequently referenced to as 24 hr operational forecasts.

In order to define an operational forecast, at UTC midnight, our pipeline takes all the Model Outlooks from the previous 24 hr, and if one or more of these outlooks indicate flaring, then the operational forecast will be that flaring will occur on this day.[4] The Model Outlook and operational forecasts are then discussed in Sections 4 and 5 for the validation data. For the holdout set, the model is retrained using all of the training and validation sets with all the hyperparameter and threshold values determined during the initial training phase. This fully trained model is then evaluated in Section 5.1.

[4] It is possible that a stricter criterion may perform better; however, this would require further thresholding and optimization, while the intention here is to provide a proof-of-concept on how `ARTop` data can be utilized with ML. As such, a more in-depth definition for this alert will follow in another publication.

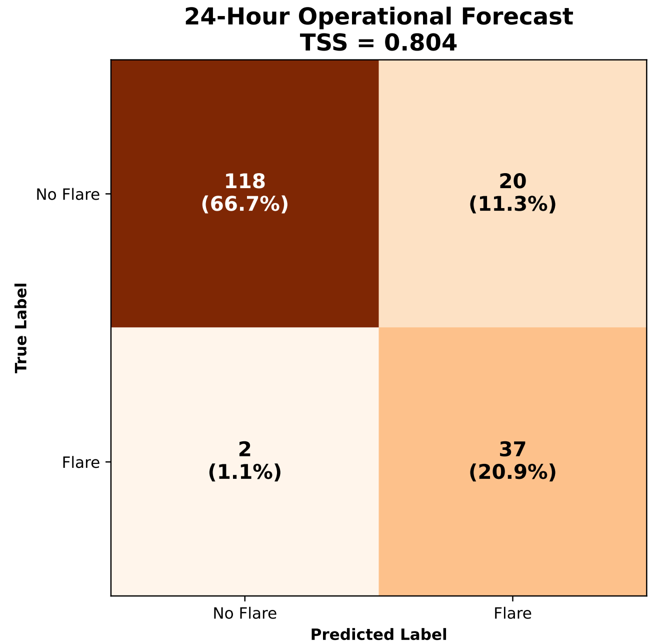

Figure 3. Confusion matrix for the `XGBoost` operational forecasts for all SHARP data in the validation set. The TSS score is shown for illustration.

4. Model Evaluation

Upon completion of the 10 cross-validation folds for the hyperparameter tuning and final model training, the model performance is evaluated for the validation set. The TSS score of 0.804 obtained for the operational forecasts indicates the model has developed good skill at accounting for class imbalance. Here, the model's hit rate exceeds the false alarm rate by 80.4% with detections predominantly being true positives. The TSS score also indicates that alongside strong discrimination of events, the model is able to meaningfully extract signal from minority cases (in this case, flares) and that the detection threshold has been well optimized. This is further evidenced when we analyze the full 720 s cadence data (Model Outlook). From this, accurate outcomes are captured by the model for 92.2% of the 21,272 observations that form the validation set.

A further breakdown of the classifier events is given in Figure 3 for the confusion matrix of the operational forecast for the validation set. When the Model Outlook is sampled to a 24 hr operational forecast, i.e., the cadence that would be used for a deployed live predictor, the accuracy decreases from 92.2% but remains high at 87.6%, likely due to an FP rate of 11.3%. These FP outcomes may be skewed slightly due to the decreased granularity of the dataset when deployed as operational forecasts, which is explored further in Section 5.

In Figure 4, SHapley Additive exPlanations (SHAP) analysis (S. M. Lundberg & S.-I. Lee 2017) summary plots are presented for the 20 most important features in the model results for the training (left) and validation (right) sets. The sign of the SHAP value (ϕ) indicates whether a positive outcome prediction, i.e., a flare, is more (positive ϕ) or less

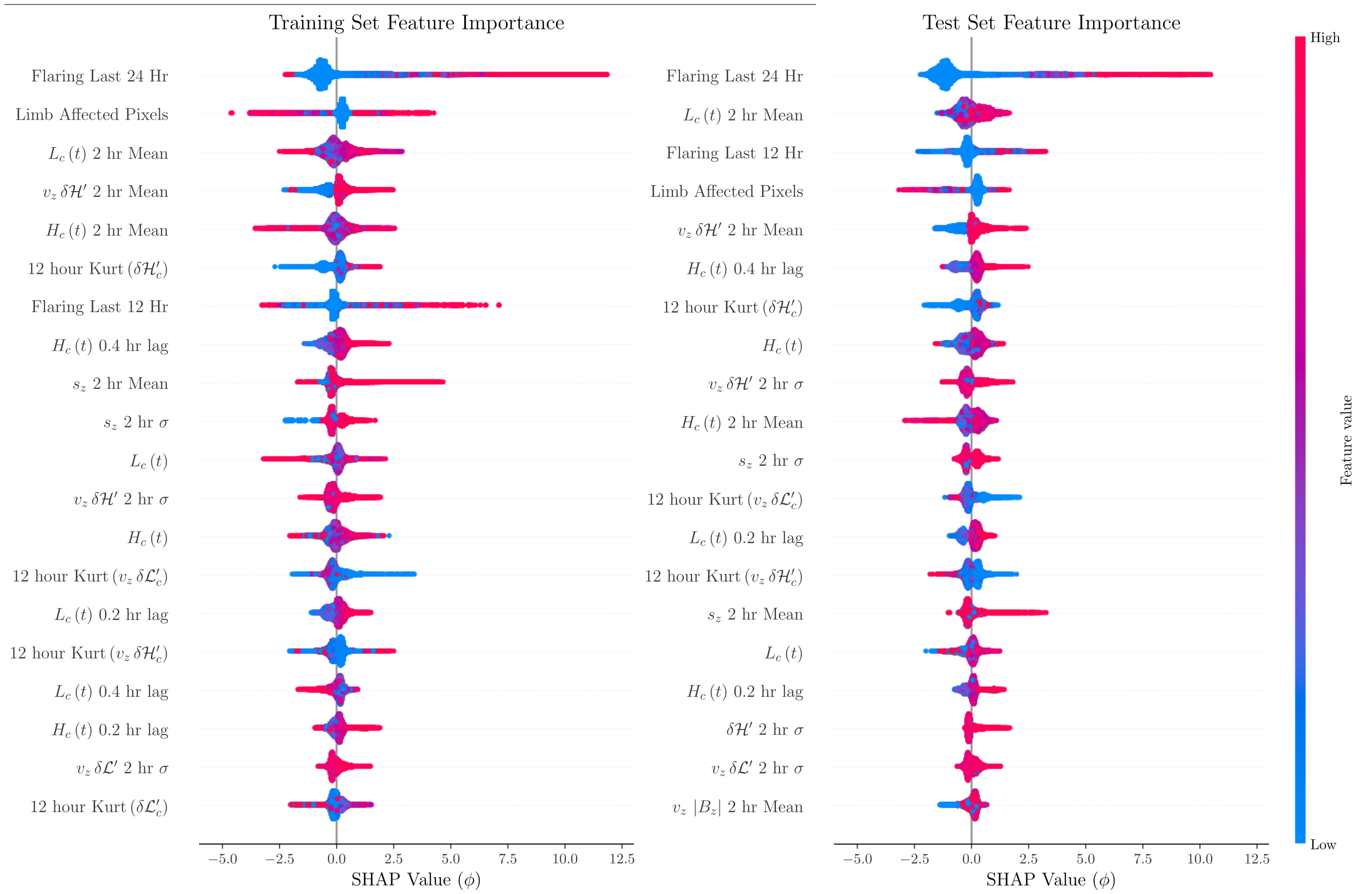

Figure 4. SHAP analysis (S. M. Lundberg & S.-I. Lee 2017) summary plots of the `XGBoost` classification model for the Training (left) and Validation (right) sets. The top 20 most important features are shown in descending order, where the feature value increases from low (blue) to high (red). This indicates how each value of the features positively/negatively impacts the magnitude of the prediction by increasing/decreasing the SHAP value, ϕ.

(negative ϕ) likely. The feature values increase from low (blue) to high (red), which along with their respective position on the *x*-axis, indicate the degree to which each feature impacts ϕ. In the two datasets, there are relatively consistent feature rankings, which implies that there is good generalization of the model. As expected, there is some deviance between the feature ordering and pattern in the training and validation sets. This indicates that first, the model has been appropriately tuned to prioritize physically plausible features for flaring. Second, these results suggest that the model has a good chance of being robust on unseen data across different periods of the solar cycle.

It is evident that the most prominent feature in the model is the flaring history, with the previous 24 hr history taking top spot in both datasets, while also providing the largest range in ϕ. In both sets, large clusterings of low values are seen to negatively impact ϕ, indicating that previously nonflaring regions have a strong weighting to remain nonflaring in the predictions made by the model. Typical values of the accumulated winding and helicity ($\mathcal{L}_c$ and $\mathcal{H}_c$) over the previous 2 hr have minimal impact on ϕ with a large cluster near 0. However, when $\mathcal{L}_c$ and $\mathcal{H}_c$ are large, they can also lead to large positive and negative ϕ values in the training set, indicating that decision trees are combining these (and other) parameters to make new metrics. Similarly to WPM, the kurtoses and other time series statistics such as the running mean and variances are capturing evolution patterns within the data that are seemingly meaningful for flare prediction.

The other quantity the authors wish to draw attention to is the limb-affected pixels. As has been previously mentioned, this feature gives the percentage of the number of pixels that exceed longitudes of $\pm 60^{\circ}$ relative to the central meridian. In both the training and validation sets, the low values (i.e., when no projection effects are present) only marginally increase ϕ, while large values (i.e., when there are projection effects present) have large impacts on the predicted outcome, for both positive (flaring) and negative (nonflaring) outcomes. This suggests that the `XGBoost` algorithm is making connections between the behavior of topological quantities and how much of the SHARP is subjected to projection effects. This will be discussed in more detail later.

5. Results and Analysis

In this section, we will first present the results from the validation set before proceeding with the holdout set. In Figure 5 the flare predictions are presented for five flaring regions, while the nonflaring regions are shown in Figures 6 and 7. In these three figures, the Model Outlook yes/no predictions are shown every 720 s about the likelihood of flaring within the next 24 hr immediately following that SHARP observation time. These outcomes are then collated into 24 hr periods spanning midnight to midnight for each day the SHARP region is present. Within each of these 24 hr periods, if the Model Outlook has a positive prediction for one

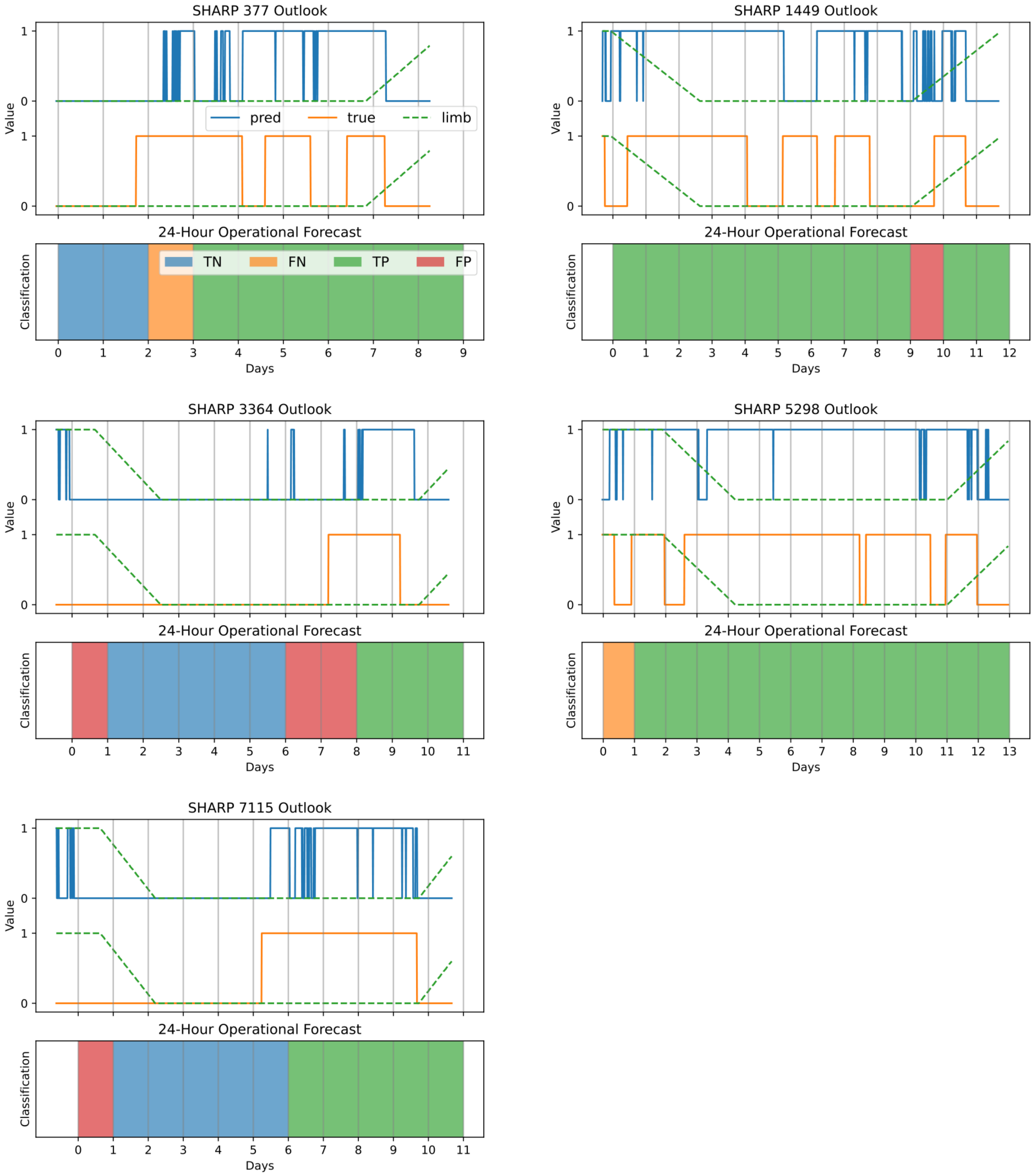

Figure 5. The Model Outlook and operational forecasts are provided for the flaring regions in the validation set. Here, the outlook provides a yes/no outcome for flaring in the following 24 hr from each SHARP observation. The ground truth is shown in orange, while the model prediction is shown in blue. The percentage of limb-affected pixels for the SHARP region is indicated in dashed green. The corresponding operational forecast shown below the Model Outlook collates these 720 s cadence predictions, and if a positive outcome exists within a 24 hr period (midnight to midnight), determines that a flare is likely the next day. Here, true negative, false negative, true positive, and false positive are indicated in blue, red, green, and orange, respectively, while the boundary between days is shown by light gray vertical lines.

of the observation periods, the operational forecast determines that a flare is the likely outcome for the following day.

Across the five regions shown in Figure 5, the operational forecast makes correct predictions (TN + TP) for flaring outcomes the following day for 87.5% of the days. Here, two FN outcomes are made, both at the onset of flaring, for SHARPs 377 and 5298. For SHARP 5298, this occurs in a period where the entirety of the SHARP data is at longitudes beyond $-60°$. Furthermore, as can be seen in Figure 8, only 4 data points exist for this initial prediction and so accumulated and running metrics employed by the model have not yet matured for the region. For SHARP 377, Figure 8 shows that the SHAP summary plots for the top 10 metrics indicate that they all mostly negatively impact ϕ, which leads to negative (nonflaring) predictions. Interestingly, the 12 hr running kurtosis for $v_z\ \delta\mathcal{L}_c{}'$ all positively impact ϕ, suggesting that a change in the active region's magnetic winding has been observed that forewarns of large flares.

In a similar vein to SHARP 5298, false positive detections are seen for SHARPs 3364 and 7115 at the start of their observation periods, where they are affected by projection effects. For SHARP 7115, the onset of flaring is then accurately captured, something that was shown to be difficult for an anonymous model in a collaborative workshop (S.-H. Park et al. 2020). For SHARP 3364 between days 6 and 8, we can see there are a few anomalous spikes in the Model

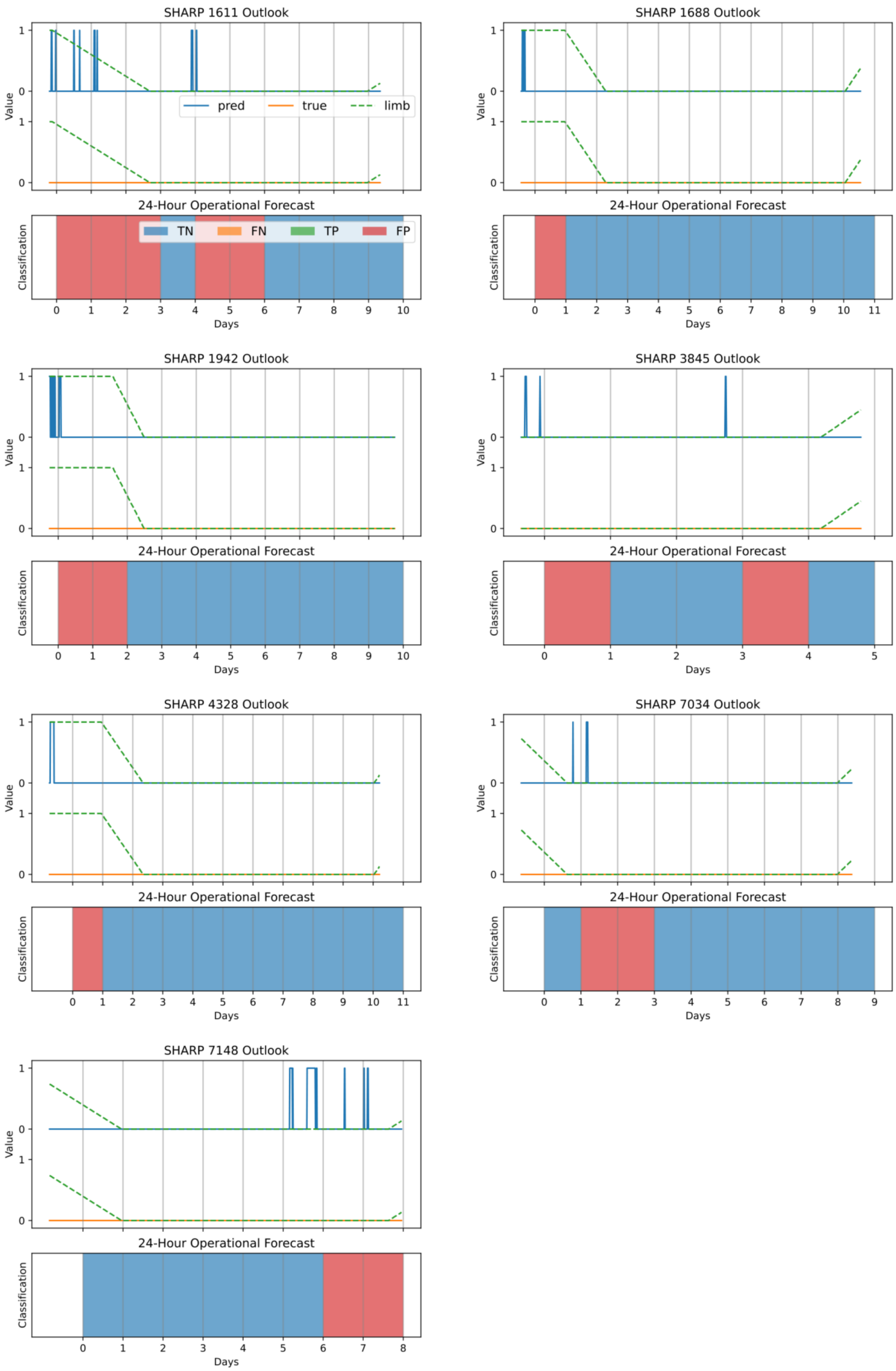

Figure 6. Same as Figure 5 but for the nonflaring regions in the validation set that have FP alerts.

Outlook prior to the onset of flaring that lead to FP predictions in the operational forecast for the 2 days preceding M/X-class flaring. The subsequent SHAP analysis for these FP periods (Figure 9) indicates that—like with SHARP 1449's FP prediction for day 9—the majority of the metrics contribute to a positive ϕ. Unlike SHARP 377, where the 12 hr running Kurt($v_z\ \delta\mathcal{L}_c{}'$) correctly indicates flaring, for SHARP 3364, it is the main contributor to the FP prediction alongside the 12 hr Kurt($v_z\ \delta\mathcal{H}_c{}'$). We also note that similar results are seen for the FP SHAP analysis of SHARP 3364 during the limb-affected

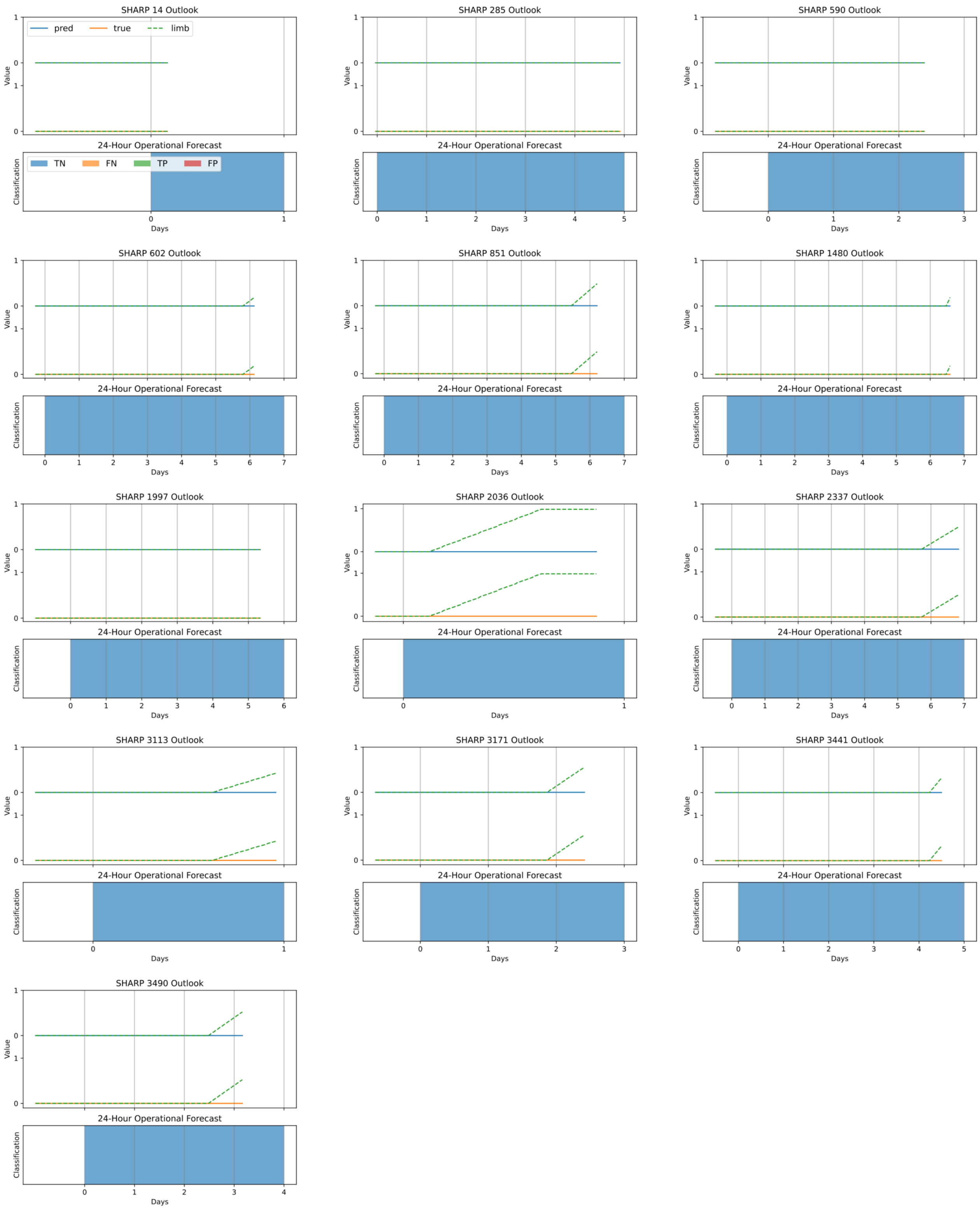

Figure 7. Same as Figure 5 but for the nonflaring regions in the validation set that have only true negative outcomes.

period of days 0–1 (not shown) as what is seen during days 6–8 in Figure 9.

As for the predictions made for the nonflaring regions, these are segregated into regions that contain TN + FP (Figure 6) and those that only contain TN (Figure 7) predictions. Overall, the model performs remarkably well, with a prediction accuracy of 87.6%, i.e., 87.6% of the predictions made correspond to either TN or TP outcomes when compared to the known flaring history of the regions analyzed.

As with the flaring regions, many of the FP predictions made for the nonflaring regions correspond to periods where projection effects are present, with the most notable exception being SHARP 7148. Once again employing SHAP analysis for these FP periods (Figure 9), it is evident that the 12 hr running

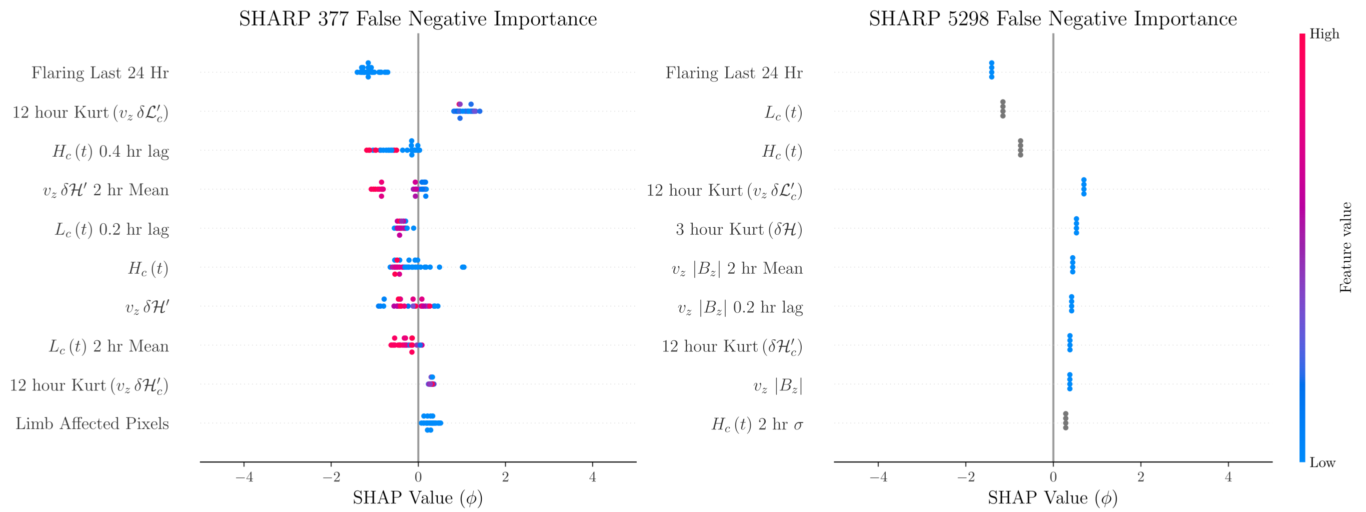

Figure 8. SHAP summary plots for the top 10 features during the periods in the Model Outlook that lead to false negative predictions for SHARP regions 377 and 5298.

kurtoses for $v_z\ \delta\mathcal{L}_c{}'$ and $v_z\ \delta\mathcal{H}_c{}'$, alongside other running and accumulated helicity-/winding-based metrics, lead to the FP predictions. This, alongside the insights gleaned from SHAP analysis on SHARPs 1449 and 3364, indicates that perhaps more nuance is required for the interpretation of the changing behavior of the helicity and winding when it comes to flare prediction. This may be that information on the type of region or the period during the solar cycle the region exists are required alongside the metrics employed in this study. These pieces of information could be incorporated into the `XGBoost` model, or similarly to the work of S. Guastavino et al. (2022, 2025), could be implemented in parallel to our `XGBoost` model, where a top-level algorithm makes a prediction based upon all the data available from the parallel models.

5.1. Holdout Set

Following the previous work of this section, the model is retrained to incorporate the validation and training sets so as to be used for deployment. Here, the hyperparameters found during the initial tuning are adopted so as to remain consistent for analysis on a completely unseen holdout set (Table 2) of data.[5]

The deployable version of our model is then tested on the holdout set. As with the training and validation sets, we also perform SHAP analysis on the retrained model applied to the holdout set (Figure 10). The top 20 parameters ranked by importance show some minor shuffling in rank, which as with the validation and training set comparison, is to be expected due to the different data employed. However, the general trends denoted previously are still present in the fully trained model on the fully unseen data further indicating that meaningful trends have been learnt during training when it comes to flare prediction.

The confusion matrices for this data are shown in Figure 11 for the operational forecast. Compared to the validation set, we find the TSS score now decreases from 0.804 to 0.524 for the operational forecast. For the full dataset (left panel in Figure 11), the FP rate has increased 13% while the TP outcome has decreased 10% for this holdout dataset. These FP periods can be visualized in Figure 12. As we see with the validation set, the limb-affected periods are also leading to FP predictions for the operational forecast with the holdout set. If these limb-affected periods are omitted from the evaluation (but remain included in the training), the fully trained model yields Model Outlook and operational forecast accuracies of 87.7% and 76.9% compared to 85.2% and 72.6%, respectively, for the full data. The TSS scores show no change when the full data (0.524) or disk-only data (0.521) are analyzed for the operational forecast. However, visual inspection of the Model Outlook and operational forecast predictions in Figure 12 show that the model performs similarly to the validation set when it is not impacted by projection effects, though there is a greater tendency for FP outcomes due to incorrect short-duration positive predictions in the Model Outlook. This could likely be mitigated by employing temporal importance to these short-duration predictions, such as defining a criterion where X positive predictions within a certain period are required for them to be counted. Additionally, it is possible that assigning some form of SES on the 24 hr window to weight the importance of more recent predictions more heavily than those nearer the start of the prediction window may also mitigate these occurrences, though this is beyond the scope of this initial study.

Encouragingly, the operational forecasts for SHARPs 11968, 12392, and 12404 (Figure 12) accurately capture the start of the flaring phase, and in the case of SHARP 11968, it also accurately determines when flaring desists, though this does occur after two FP days due to limb-affected data. For SHARP 12404, an FN outcome exists on the final flaring day, though it is worth noting that this period occurs when the region transitions from longitudes $>60^\circ$ to $<60^\circ$. For SHARP 12392, we obtain FP outcomes once flaring has stopped, though it is worth noting that the active region is subjected to numerous C-class flares during this period, with C3.7 and C4.0 flares occurring within 0.75 hr of each other on day 8. Similarly, SHARP 7357 has several larger C-class flares, with the most notable ones occurring on day 2 being C9.9 and C7.3.

[5] We note that while the validation set is not used to tune hyperparameters nor during the initial training, it is used to define and tune the probability threshold for what constitutes a detection, and so an additional holdout set is required for assessing true unseen data.

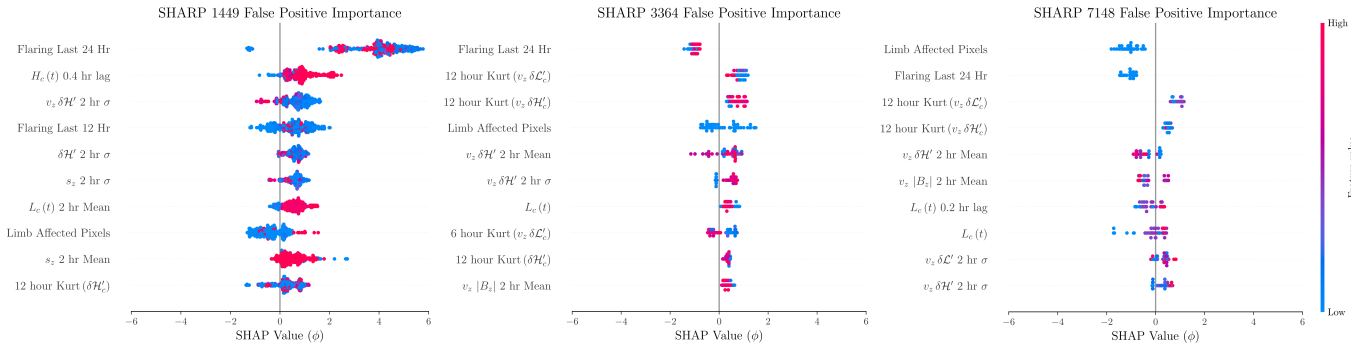

Figure 9. SHAP summary plots for the top 10 features during the periods in the Model Outlook that lead to false positive predictions for SHARP regions 1449, 3364 (days 6–8), and 7148.

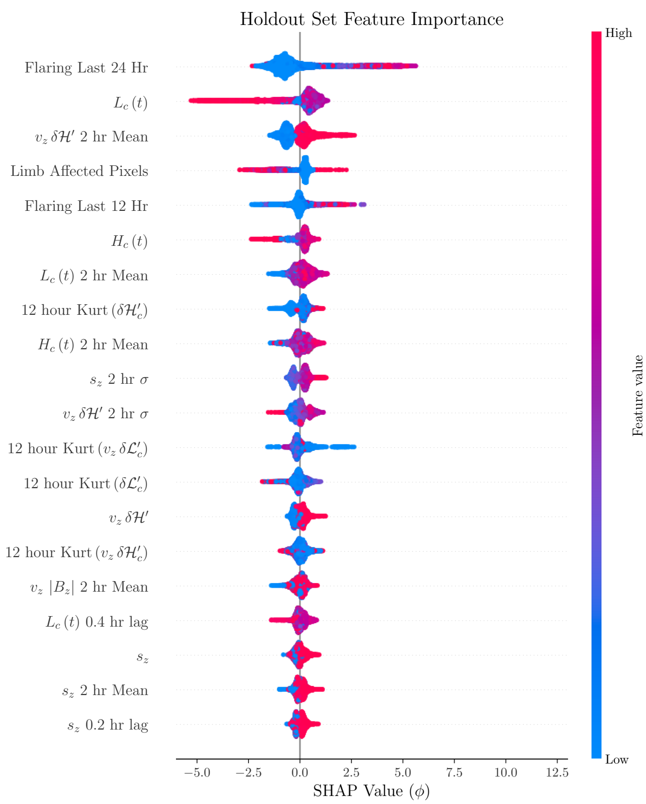

Figure 10. Same as Figure 4 but for all the regions in the holdout set.

In combination with other small flares, the flaring history over the previous 12 and 24 hr would indicate that a significant accumulation of topology is being released and that the region may be considered as flaring. As can be seen in the SHAP analysis for the model training (Figure 4), the previous flare activity is the parameter with the greatest weighting for the model and so the cumulation of these smaller flares may be wrongly influencing the model prediction in instances where numerous C-class flares are observed.

According to Table 2, SHARPs 4071 and 12405 both emitted two M-class flares and a number of C-class flares. However, in Figure 12 it is evident that no flaring occurs for either region. This is because these flares did not occur until the active regions rotated beyond the view of SDO, though it is worth mentioning that they were actively emitting several larger C-class flares when visible with SDO. From day 5 onward, the operational forecast for SHARP 4071 yields only FP outcomes (Figure 12), with some of these alerts being the cause of short-duration spikes in the Model Outlook, particularly around days 4–6, with a longer spike starting on day 6 that persists into day 7. This longer duration prediction may be subjected to projection effects as the limb-affected pixels do become nonzero toward the end of this period. Furthermore, during these FP outcomes, the active region exhibits numerous C-class flares, with one 5.5 hr period consisting of C5.3, C2.0, C2.6, C1.8, and a C1.1 flare, while another day sees a C9.2 flare. Similarly for SHARP 12405, C6.0 and C4.2 flares occur within 20 minutes of one another on day 5, while day 6 sees an accumulation of C3.4, C2.4, and C1.9 flares. There is no further activity until the very end of day 9 when the first M-class flare (M1.6) erupts. As with SHARP 12392, the steady release of magnetic energy through more frequent, smaller flares for these two active regions will have an impact on the previous flaring activity parameters of the model, which could be the cause of these FP predictions.

Furthermore, it is likely that the model requires refinement by training with more regions where the maximum flare magnitudes are M-class flares. The results presented here are a promising first step in building a live predictor based on `ARTop` calculations, but it is evident that regions with numerous C-class and smaller M-class flares are currently problematic for the deployed model. Similarly, further improvement is also needed when processing "near-limb" data. As is indicated in the SHAP analysis, the limb-affected pixels parameter does appear to have important weighting within the `XGBoost` model, though given the results seen for the holdout and validation sets, this alone is unable to mitigate the manner in which topological quantities like the winding and helicity are affected.

6. Concluding Remarks

This manuscript presents a new solar flare prediction model based upon the magnetic topology calculations of `ARTop` that has been trained on 232 SHARP regions and assessed on a further 12 regions. To ensure that overlapping windows are avoided during flare prediction—and subsequently the duplication of data introducing bias into those predictions—the Model Outlook is collated into distinct prediction windows of a UT day. During the training phase of the model using these distinct prediction windows, a favorable TSS score of 0.804 is obtained on the validation set. The subsequent SHAP analysis performed on the training and validation sets highlights that the model has developed good generalization, indicating that it has learnt meaningful trends in the underlying data and how

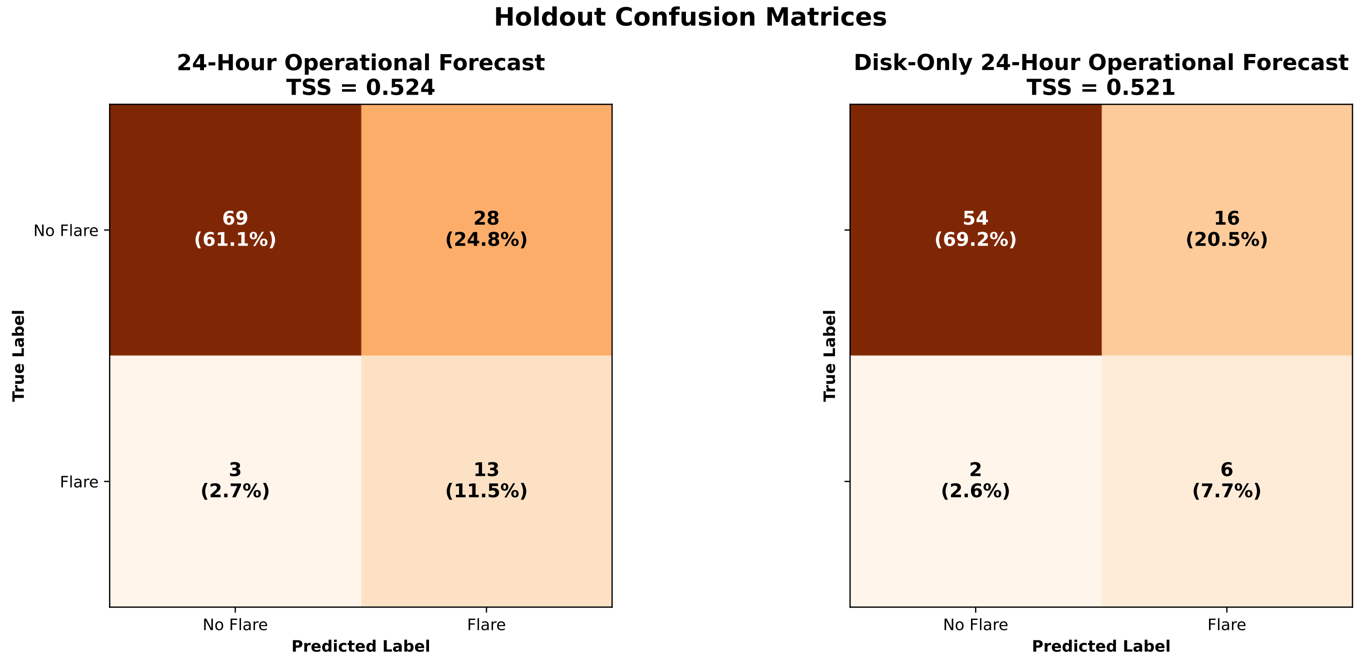

Figure 11. Confusion matrix comparison plots for the `XGBoost` operational forecasts for all SHARP data in the holdout set. The left panel provides results for the full dataset, while the right panel omits limb-affected data. True Skill Statistic scores are shown for illustration.

they relate to flaring that exceed magnitudes of M1.0 in the next 24 hr. On the 12 regions that form the holdout set, the TSS score (0.524) does decrease but the accuracy (72.6%) remains high for the daily operational forecast predictions. This is despite a number of the regions being problematic due to projection effects and frequent C-class flares resulting in the release of stored magnetic energy.

Currently, direct comparisons are difficult between competing flare prediction models—indicating the need for standardized data, such as that outlined in A. Hollanda et al. (2021), to be employed across studies. For example, M. G. Bobra & S. Couvidat (2015) and T. Cinto et al. (2020) compare model performance metrics across several studies on flare prediction. In these studies, it is shown that most models typically have TSS scores ranging between 0.3 and 0.7, but as is highlighted in M. G. Bobra & S. Couvidat (2015), Table 2, the intervals between predictions, the data used for evaluation, and training all differ. When M. G. Bobra & S. Couvidat (2015) tune their model for TSS, they obtain a score of 0.761 when using an operational model (24 hr prediction intervals). Similarly, while N. Nishizuka et al. (2017) find TSS scores above 0.9 for their model, outperforming the NICT space weather forecast center (TSS = 0.21), this calculation is only for 29 X-class flares and did not include limb-affected data/statistics. On the other hand, the CNN model of C. Pandey et al. (2023) yields a TSS score of 0.54 for flares exceeding M1.0 within the next 24 hr. Similarly, Deep Flare Net (DeFN; N. Nishizuka et al. 2021) yields TSS scores of 0.80 for flares M1.0 and above and 0.63 for flares C1.0 and above. However, their operational assessment in 2019–2020 saw scores decrease to 0.24(0.48) for M-class flares when a probability threshold of 50(40)% is adopted.

By comparison, the work presented in this manuscript explores data that contains a total of 2142 flares, of which there are 36 X-class and 384 M-class flares. When evaluating the training model with the validation set, we obtain a TSS score of 0.804 for the operational forecasts. However, when employing the fully trained model on the holdout set, the score decreases by 0.524, respectively. If the limb-affected data are ignored, such as with N. Nishizuka et al. (2017), then the TSS scores remain similar for our deployable model (0.521), showing similar predictive efficacy to DeFN. It is worth noting that unlike many of the models discussed here, the operational forecast and Model Outlook have not been optimized on TSS score but on the weighted combination of statistics given in Equation (10).

One of the main issues encountered with the fully trained model evaluation is the digital nature of what constitutes a positive prediction. That is, a hard threshold is set where an outcome is only positive (flaring) if the GOES flux $\geqslant$M1.0 for a single flare. In SHARP 4071 we witness numerous C-class flares during a 5.5 hr period where the total magnitude exceeds M1.0, which due to this digital cutoff, result in FP outcomes. Similarly, SHARP 7357 exhibits a C9.9 flare during a period where the operational forecast also yields an FP outcome. One could argue that the two scenarios encountered here, while giving the wrong or undesired outcomes, are not incorrect results but merely artifacts due to the deficiencies in how the solar flare forecasting community defines what a meaningful flare is for space weather.

The other challenge for the model presented here is posed by regions that rotate into view. First, these regions have likely undergone a significant portion, if not all, of their flux emergence phase during which `ARTop` will not have captured the buildup of topology ($\mathcal{H}_c{}'$ and $\mathcal{L}_c{}'$), which as is evidenced in our SHAP analysis, forms an important metric in the model's ability to predict flaring. This could perhaps be mitigated by estimating the topology buildup through extrapolations (e.g., R. Jarolim et al. 2023). The second, and more important issue of regions rotating into view, is the projection effects. While the limb-affected pixels parameter does appear to have important weighting within the `XGBoost` model, this metric alone is unable to mitigate the manner in which topological quantities like the winding and helicity are affected.

Before a topological based live predictor can be used for operational forecasting of solar flares, these issues must be

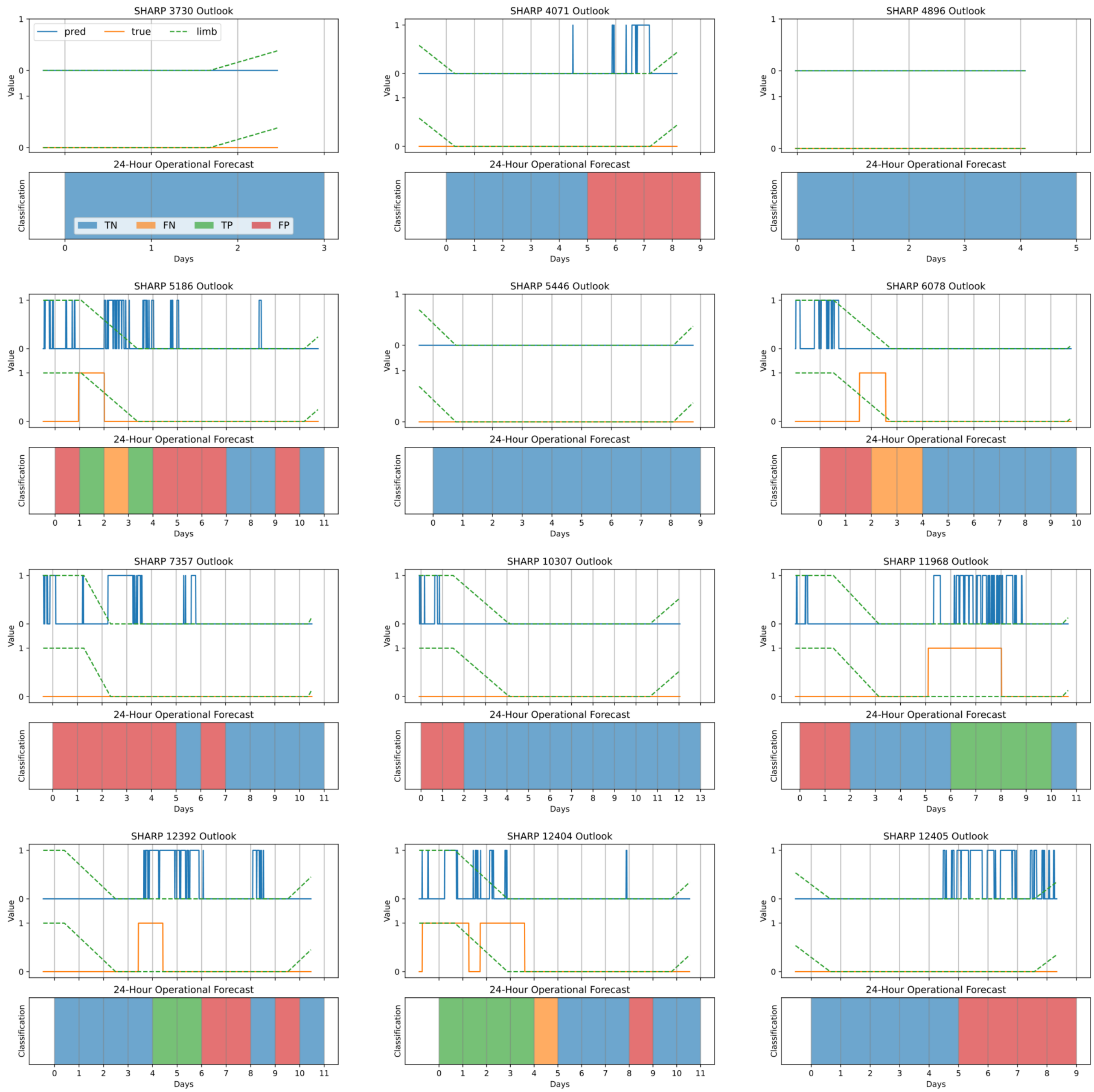

Figure 12. Same as Figure 5 but for the regions in the holdout set.

addressed. In future work, we aim to resolve the issues presented in this manuscript by continuing to update the living dataset first created by WPM, including extrapolations to approximate the topology buildup of active regions that rotate into view, alongside the incorporation of multilayered models such as those discussed in S. Guastavino et al. (2022, 2025). In light of the issues around larger/numerous C-class flares, we will also explore expanding the binary classification of flare/no flare to a multiclass classification in a similar manner to that explored in J. Bringewald & O. Parisot (2025).

Acknowledgments

The authors acknowledge support from the US Air Force grant FA8655-23-1-7247. D.M. acknowledges support from a Leverhulme Trust grant (RPG-2023-182), a Science and Technology Facilities Council (STFC) grant (ST/Y001672/1), and a Personal Fellowship from the Royal Society of Edinburgh (ID:4282). C.P. acknowledges support from the UK Science and Technology Funding Council under grant No. ST/W00108X/1. This research utilizes version 5.1.2 of the `SunPy` open-source software package (S. J. Mumford et al. 2024).

Data Availability

Research data supporting this study are openly available from Github at https://github.com/DrTomWilliams/Machine-Learning-Flare-Dataset

Research code supporting this study are openly available from Github at https://github.com/DavidMacT/ARTop.

Appendix

Here, the full list of 207 SHARP regions used for the training of the model is given in Table 3. A breakdown on the largest flare size, as well as the total number of X-, M-, and C-class flares are given for each region. In total, 57 regions exhibit flaring above M1.0, with a further 175 regions being flare-free.

Table 3
SHARP Regions Investigated with Flare Information Provided by HEK

NOAA Active Region	SHARP Number	Largest Flare	# X-class Flares	# M-class Flares	# C-class Flares
11069	8	M1.2	⋯	1	7
11071	17	⋯	⋯	⋯	⋯
11072	26	⋯	⋯	⋯	⋯
11076	43	⋯	⋯	⋯	⋯
11073	45	⋯	⋯	⋯	⋯
11079	49	M1.0	⋯	1	⋯
11080	51	C1.2	⋯	⋯	1
11081	54	M2.0	⋯	1	5
11086	83	⋯	⋯	⋯	⋯
11087	86	C3.4	⋯	⋯	5

(This table is available in its entirety in machine-readable form in the online article.)

ORCID iDs

Thomas Williams https://orcid.org/0000-0002-2006-6096
Christopher B. Prior https://orcid.org/0000-0003-4015-5106
David MacTaggart https://orcid.org/0000-0003-2297-9312
O. P. M. Aslam https://orcid.org/0000-0001-9521-3874
D. Shaun Bloomfield https://orcid.org/0000-0002-4183-9895

References

Abramenko, V. 2005, ApJ, 629, 1141
Alielden, K., MacTaggart, D., Ming, Q., Prior, C., & Raphaldini, B. 2023, RASTI, 2, 398
Antalova, A. 1996, CoSka, 26, 98
Aslam, O., MacTaggart, D., Williams, T., Fletcher, L., & Romano, P. 2024, MNRAS, 534, 444
Benz, A. O. 2017, LRSP, 14, 2
Bobra, M. G., & Couvidat, S. 2015, ApJ, 798, 135
Bobra, M. G., Sun, X., Hoeksema, J. T., et al. 2014, SoPh, 289, 3549
Boers, N. 2018, NatCo, 9, 2556
Bringewald, J., & Parisot, O. 2025, Astro, 4, 23
Brown, R. G. 1959, Journal of the Royal Statistical Society. Series A (General), 123, 348
Chen, L., Liu, R., Liu, Z.-P., Li, M., & Aihara, K. 2012, NatSR, 2, 342
Chen, T., & Guestrin, C. 2016, arXiv:1603.02754
Cinto, T., Gradvohl, A. L. S., Coelho, G. P., & da Silva, A. E. A. 2020, SoPh, 295, 93
Dakos, V., Matthews, B., Hendry, A. P., et al. 2019, NatEE, 3, 355
Deshmukh, V., Baskar, S., Berger, T. E., Bradley, E., & Meiss, J. D. 2023, A&A, 674, A159
Florios, K., Kontogiannis, I., Park, S.-H., et al. 2018, SoPh, 293, 28
Gopalswamy, N. 2022, Atmos, 13, 1781
Guastavino, S., Legnaro, E., Massone, A. M., & Piana, M. 2025, ApJ, 985, 53
Guastavino, S., Marchetti, F., Benvenuto, F., Campi, C., & Piana, M. 2022, A&A, 662, A105
Hoeksema, J. T., Liu, Y., Hayashi, K., et al. 2014, SoPh, 289, 3483
Hollanda, A., Antunes Da Silva, A. E., & Cinto, T. 2021, DIB, 37, 107203
Holt, C. C. 2004, Int. J. Forecast., 20, 5
Hurlburt, N. E. 2022, AGUFM, 2022, SH32F-1818
Jarolim, R., Thalmann, J., Veronig, A., & Podladchikova, T. 2023, NatAs, 7, 1171
Klein, K.-L., & Dalla, S. 2017, SSRv, 212, 1107
Korsós, M., Erdélyi, R., Huang, X., & Morgan, H. 2022, ApJ, 933, 66
Li, M., Cui, Y., & Luo, B. 2025, ChJSS, 45, 82
Liu, Y., Welsch, B. T., Valori, G., et al. 2023, ApJ, 942, 27
Lundberg, S. M., & Lee, S.-I. 2017, in Proc. 31st Int. Conf. on Neural Information Processing Systems (Curran Associates Inc.), 4768
MacTaggart, D., Prior, C., Raphaldini, B., Romano, P., & Guglielmino, S. L. 2021, NatCo, 12, 6621
Mumford, S. J., Freij, N., Stansby, D., et al. 2024, SunPy, v5.1.1, Zenodo, doi:10.5281/zenodo.591887
Nishizuka, N., Kubo, Y., Sugiura, K., Den, M., & Ishii, M. 2021, EP&S, 73, 64
Nishizuka, N., Sugiura, K., Kubo, Y., et al. 2017, ApJ, 835, 156
Pandey, C., Ji, A., Angryk, R. A., & Aydin, B. 2023, in IEEE Sixth Int. Conf. on Artificial Intelligence and Knowledge Engineering (IEEE), 83
Park, S.-h., Chae, J., & Wang, H. 2010, ApJ, 718, 43
Park, S.-H., Lee, J., Choe, G.-S., et al. 2008, ApJ, 686, 1397
Park, S.-H., Leka, K. D., Kusano, K., et al. 2020, ApJ, 890, 124
Pedregosa, F., Varoquaux, G., Gramfort, A., et al. 2011, JMLR, 12, 2825
Pevtsov, A. A., Maleev, V. M., & Longcope, D. W. 2003, ApJ, 593, 1217
Prior, C., & MacTaggart, D. 2020, RSPSA, 476, 20200483
Raphaldini, B., Prior, C. B., & MacTaggart, D. 2022, ApJ, 927, 156
Schou, J., Scherrer, P. H., Bush, R. I., et al. 2012, SoPh, 275, 229
Schuck, P. W. 2008, ApJ, 683, 1134
Sevim, C., Oztekin, A., Bali, O., Gumus, S., & Guresen, E. 2014, Eur. J. Oper. Res., 237, 1095
Sinha, S., Gupta, O., Singh, V., et al. 2022, ApJ, 935, 45
Temmer, M. 2021, LRSP, 18, 4
van der Sande, K., & Muñoz-Jaramillo, A. 2025, ApJ, 984, 87
Vemareddy, P. 2021, MNRAS, 507, 6037
Williams, T., Prior, C. B., & MacTaggart, D. 2025, ApJ, 980, 102
Winters, P. R. 1960, Manage. Sci., 6, 324
Yi, K., & Moon, Y.-J. 2022, AGUFM, 2022, NG52A–0178
Yun, J., & Shin, J. 2024, arXiv:2405.14750