Space Weather

RESEARCH ARTICLE
10.1029/2025SW004546

Special Collection:
The Space Weather Research to Operation to Research (R2O2R) Pipeline(s): Progress, Challenges and Prospects

Verification of the NOAA Space Weather Prediction Center Solar Flare Forecast (1998–2024)

Enrico Camporeale[1,2] and **Thomas E. Berger**[3]

[1]School of Physical and Chemical Sciences, Queen Mary University of London, London, UK, [2]Space Weather Technology, Research and Education Center, University of Colorado, Boulder, CO, USA, [3]National Center for Atmospheric Research, High Altitude Observatory, Boulder, CO, USA

Key Points:

- Space Weather Prediction Center (SWPC) solar flare forecasts do not outperform zero-cost baselines such as persistence and climatology over a 26-year validation period
- SWPC solar flare forecasts are poorly calibrated and produce excessive false alarms, particularly for X-class and in "all-clear" scenarios
- We recommend the operational adoption of modern machine learning methods and routine verification of forecast skill

Supporting Information:

Supporting Information may be found in the online version of this article.

Correspondence to:

E. Camporeale,
enrico.camporeale@colorado.edu

Citation:

Camporeale, E., & Berger, T. E. (2025). Verification of the NOAA Space Weather Prediction Center solar flare forecast (1998–2024). *Space Weather*, *23*, e2025SW004546. https://doi.org/10.1029/2025SW004546

Received 21 MAY 2025
Accepted 10 SEP 2025

Author Contributions:

Conceptualization: Enrico Camporeale
Methodology: Enrico Camporeale
Software: Enrico Camporeale
Writing – original draft: Enrico Camporeale
Writing – review & editing: Thomas E. Berger

Abstract The NOAA Space Weather Prediction Center (SWPC) issues the official U.S. government forecast for M-class and X-class solar flares, yet the skill of these forecasts has never been comprehensively verified. In this study, we evaluate the SWPC probabilistic flare forecasts over a 26-year period (1998–2024), comparing them to several zero-cost and statistical baselines including persistence, climatology, Naive Bayes, and logistic regression. We find that the SWPC model does not outperform these baselines across key classification and probabilistic metrics and exhibits severe calibration issues and high false alarm rates, especially in high-stakes scenarios such as detecting the first flare after extended quiet periods. These findings demonstrate the need for more accurate and reliable eruption forecasting models which we suggest should be based on modern data-driven methods. *The findings also provide a standard against which any proposed eruption prediction system should be compared. We suggest that space weather forecasters regularly update and publish analyses like the one demonstrated here to provide up-to-date standards of accuracy and reliability against which to compare new methods.*

Plain Language Summary The U.S. government regularly provides space weather forecasts, including predictions of X-rays from solar eruptions—sudden bursts of energy from the Sun that can harm satellites, power grids, and astronauts. Many people and organizations use these official forecasts when planning activities that depend on space weather. In this study, we looked closely at how accurate those forecasts have been over the last 26 years. We compared them to very simple methods, like just assuming tomorrow will be similar to today, or using past averages. Surprisingly, we found that the official forecasts were no better than these basic methods. In some important situations—like when the Sun has been quiet for weeks and a powerful flare might occur—the forecasts often failed. This is a serious concern, especially for astronaut safety. We suggest that modern data-driven techniques, like machine learning, could produce better predictions than current techniques and should be prioritized for development. We also encourage space weather forecasters to regularly test and publish how well their predictions are working.

1. Introduction

Solar eruptions are among the most energetic phenomena in the heliosphere, capable of releasing up to 10^{32} erg of energy in the form of electromagnetic radiation, energetic particles, and plasma ejections (Benz, 2017). These events are key drivers of space weather, with the potential to disrupt satellite operations, degrade radio communications, and endanger astronauts in space (Schrijver et al., 2015; Schwenn, 2006; Temmer, 2021). Accurate forecasting of solar eruption occurrence is therefore of critical importance for operational space weather services.

The first indication of a solar eruption to reach Earth is the electromagnetic radiation generated primarily in X-ray and Extreme Ultraviolet (EUV) photons. Colloquially referred to as "solar flares", solar eruption photonic emission is classified according to its peak soft X-ray flux in the 1–8 Å band as observed by the NOAA/GOES satellites. The original flare classification system included Common (C), Medium (M), and eXtreme (X) classes. Later classifications included lower-level X-ray emission in the A and B classes. Each class represents a tenfold increase in peak 1–8 Å flux. Of particular concern for space weather are M-class and X-class flares, which can cause moderate to severe high frequency (3–30 MHz) radio and radar interference on the sunlit side of the Earth and are often associated with large plasma eruptions (so-called "Coronal Mass Ejections" or CMEs) that, if they are Earth-directed, can cause moderate to extreme geomagnetic storms (Gopalswamy, 2018; Hill et al., 2005).

Several solar flare catalogs are maintained to support research and operational activities. These include the NOAA event lists derived from GOES X-ray data, the Heliophysics Event Knowledgebase (HEK) (Hurlburt et al., 2012), and curated flare event databases such as those produced by the Solar Data Analysis Center (SDAC) and the Kanzelhöhe Solar Observatory (Chen et al., 2024; Pötzi et al., 2015). *A solar flare catalog based on NASA Solar Dynamics Observatory (SDO) Atmospheric Imaging Assembly (AIA) EUV images was recently presented in* Van der Sande et al. (2022). These catalogs are foundational resources not only for retrospective event studies but also for the development and validation of empirical and machine learning-based flare forecasting models (Camporeale, 2019; Florios et al., 2018; Georgoulis et al., 2024; Leka & Barnes, 2018; Leka et al., 2019). Recently, Berretti et al. (2025) *produced a new catalog called ASR (Archival Solar Flares), which includes a method to locate flare locations on the Sun and link them to photospheric active regions using SDO data. In this work, we used the ASR v1.0.0 (released on 12 March 2025) available on* https://github.com/helio-unitov/ASR_cat/releases/download/v1.1/f_1995_2024.csv. *The data set covers flares that occurred during the period 2002–2024.*

The official US government source of solar eruption/flare forecasts is the NOAA Space Weather Prediction Center (SWPC) in Boulder, Colorado, which provides operational (i.e., continuously available with robust backup capabilities) forecasts for M- and X-class flares. Three-day solar flare forecasts are issued every 12 hr and summarized daily (the latest 3-day forecast is available at https://www.swpc.noaa.gov/products/3-day-forecast *and summaries are available at* https://www.swpc.noaa.gov/products/report-and-forecast-solar-and-geophysical-activity). *The 3-day flare forecasts use the SWPC-specific "Radio Blackout" or R-scale for flare classification where an R1 flare corresponds to an M1–4.9 X-ray peak, R2 corresponds to an M5–9.9 flare, and R3–R5 classes correspond to X-class flares. In contrast, the summary report and forecast uses the M- and X-class nomenclature. In spite of the differing nomenclature both forecast products consist of three integer numbers that are the forecast probability of the occurrence of one or more M- or X-class flares for each of the next 3 days (see* https://www.swpc.noaa.gov/noaa-scales-explanation *for an explanation of NOAA/SWPC space weather scales). The SWPC forecast is used by a range of stakeholders, including satellite operators, airlines, air traffic controllers, radar operators, and defense agencies.* Despite their central role in space weather forecasting, the methodology behind SWPC's flare probability forecasts is not publicly documented in detail, and, to the best of the authors' knowledge, no comprehensive verification study of their accuracy and reliability, or skill, has been published *since the work of* Crown (2012), *which covered the period of 1996–2008. More recent studies, such as* Leka et al. (2019), *verified the forecasts over only a short period of time (2016–2017).*

SWPC flare forecast methodology is not described on the NOAA website, however, according to Leka et al. (2019) the methodology begins with classifying active regions using the McIntosh scheme (McIntosh, 1990) and assigning flare probabilities based on historical flaring rates established for each McIntosh class (i.e., a climatological baseline); these probabilities are then modified by human forecasters based on region evolution, recent flaring activity, and expert judgment, and the resulting region-specific probabilities are aggregated into a full-disk forecast, which may be further adjusted considering the activity of recently rotated-off or returning active regions and, when available, supplemented by additional model outputs; the initial forecast is issued at 22:00 UTC for the next day and incorporated into the official 3-day forecast released at 00:30 UTC and updated at 12:30 UTC.

The goal of this study is to systematically verify the skill of SWPC's probabilistic forecasts for M-class and X-class flares *over the period 1998–2024 (26 years of data)*. We compare forecast probabilities with flare occurrence records from NOAA's GOES-based catalog, employing a number of standard metrics for classification. This analysis contributes both to operational forecasting assessment and to the broader objective of benchmarking baseline models for flare prediction, which is increasingly important in the era of data-driven and machine learning-based space weather modeling.

2. Data

This study uses three primary data sources: the official daily probabilistic forecasts issued by NOAA's Space Weather Prediction Center (SWPC), the flare occurrence records from the NOAA event reports, and the ASR catalog.

NOAA/SWPC issues daily forecasts for the probability of at least one M-class or X-class flare occurring within the next 1, 2, and 3 days. Historical records of these forecasts, dating back to 1996, are available from

Table 1
Total Counts of Positive and Negative Days for M-Class and X-Class Flares for the Period 1998–2024 (9,828 days)

	M-class	X-class
Positive days (label = 1)	2,021	254
Negative days (label = 0)	7,807	9,574
Class imbalance ratio (neg/pos)	3.86	37.69

the SWPC data archive at ftp://ftp.swpc.noaa.gov/pub/warehouse/. These records are stored as annual text files. We parsed the forecasts from each file and compiled them into a unified CSV format for further analysis. Each row in the resulting data set corresponds to a unique date and contains the forecast probabilities for M-class and X-class flares issued on that day (six numbers: 1, 2, 3-day ahead for M- and X-class). Flare occurrence data were obtained from two sources. For the years 2002–2024, we used the ASR flare catalog, which includes flare events curated from the GOES X-ray flux records and incorporates the official NOAA event data (https://github.com/helio-unitov/ASR_cat). For the years 1996–2001, flare records were retrieved directly from the NOAA SWPC event reports, available at ftp://ftp.swpc.noaa.gov/pub/indices/events/.

To perform the validation, we constructed a binary "ground truth" label for each day in the data set. A day is labeled as positive (1) for a given class (M or X) if at least one flare of that class occurred during that day (UTC). Otherwise, it is labeled as negative (0). This daily binarization of the flare records aligns with the forecast product, which specifies probabilities of at least one event per day.

The merged data set spans 10,338 unique days. *As explained in* Section 3.5, *we do not perform a statistical analysis of the first 17 months (08-1996 to 12-1997) and instead use those data as a training buffer for baseline models. For the remaining period (1998–2024),* Table 1 summarizes the total number of positive and negative instances for both M-class and X-class flares *(once again, those are the number of days with at least one flare, the information about how many flares occur in a day is not relevant).* The distribution is highly imbalanced, particularly for X-class flares, which occur far less frequently than M-class flares.

To illustrate long-term solar cycle variability, we computed a 27-day rolling *(non-overlapping)* sum of positive flare days and plotted it as a function of time. Figure 1 shows the number of days within each 27-day window during which at least one M-class or X-class flare occurred. This smoothed activity index clearly tracks the solar cycle, with pronounced peaks near solar maxima and extended periods of low activity near solar minima *(27-day average sunspot number shown as blue dots)*. The 27-day window was chosen to approximate the solar rotation period, capturing active-region recurrence patterns.

We also analyzed the seasonal distribution of flares. Figure 2 shows the number of M-class and X-class flares as a function of the day of the year, aggregated over all years. This seasonal profile highlights the variability of flare activity, which is modulated by the solar cycle but does not show strong annual periodicity.

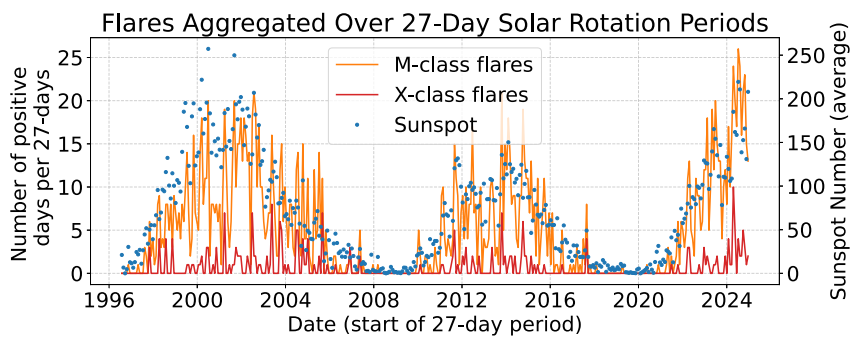

Figure 1. Number of positive days (M-class or X-class) in a 27-day sliding window, plotted over the study period. Blue dots indicate the 27-day average of sunspot numbers. Modulation due to the approximately 11-year solar magnetic activity cycle is clearly visible.

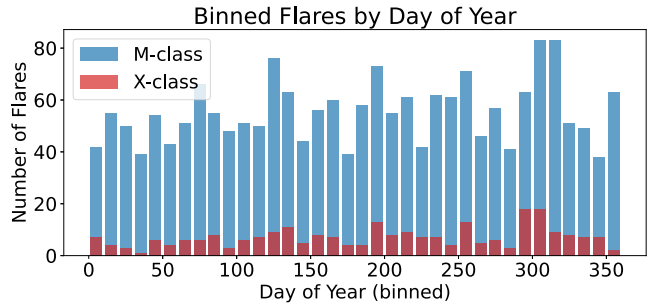

Figure 2. Total number of M-class and X-class flares as a function of day of year, aggregated over the entire data set.

In addition, we computed the conditional probability that a flare would occur as a function of the number of preceding flare-free days. This quantity is useful to assess the memory effects or persistence in flare activity. Figure 3 displays the empirical conditional probabilities for both M-class and X-class flares.

For M-class flares, the conditional probability of occurrence is greater than 50% if a flare occurred on the previous day (*zero prior flare-free days*), indicating strong short-term persistence in flare activity. In contrast, the conditional probability for X-class flares is around 20% after a flare day. For both flare classes, the conditional probability decays with increasing flare-free duration and appears to plateau after approximately 10 days, suggesting limited memory beyond that scale.

3. Methodology

To evaluate the skill of the SWPC probabilistic forecasts for M-class and X-class flares, we adopt a suite of metrics applicable to both deterministic and probabilistic forecasts. The SWPC predictions are issued as probabilities, so we convert them into binary (yes/no) predictions using a probability threshold and compute standard classification scores. In addition, we assess the forecasts directly in their probabilistic form using proper scoring rules and reliability diagrams.

3.1. Confusion Matrix and Binary Metrics

Given a threshold θ, we define a flare-day prediction if the forecast probability $p \geq \theta$. This thresholding converts the probabilistic forecasts into binary predictions, which can be compared to the ground-truth binary labels (flare or no flare) for each day.

From this comparison, we construct a confusion matrix:

- *True Positive (TP)*: the forecast predicted a flare and a flare occurred.
- *False Positive (FP)*: the forecast predicted a flare but no flare occurred.
- *True Negative (TN)*: the forecast predicted no flare and no flare occurred.
- *False Negative (FN)*: the forecast predicted no flare, but a flare occurred.

Based on these quantities, we compute the following deterministic performance metrics:

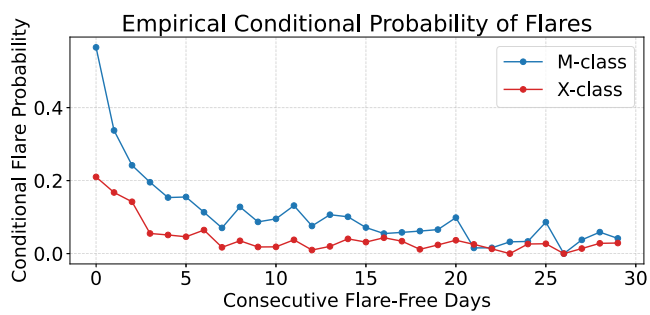

Figure 3. Conditional probability of observing an M-class or X-class flare given n consecutive prior days without such a flare.

- *Accuracy (ACC)*: the proportion of correct predictions (both flare and no-flare days):

$$\mathrm{ACC} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FP} + \mathrm{TN} + \mathrm{FN}}$$

- *Precision (PREC)*: the proportion of predicted flare days that were correct (*i.e., the likelihood that a flare occurs, when the model predicts a flare*):

$$\mathrm{PREC} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}$$

- *Recall (REC) or Probability of Detection (POD)*: the proportion of actual flare days that were correctly predicted:

$$\mathrm{REC} = \mathrm{POD} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} = \frac{\mathrm{TP}}{\mathrm{P}}$$

- *False Alarm Ratio (FAR)*: the fraction of predicted flare days that were false alarms (*i.e.*, *the likelihood that a flare does NOT occur, when the model predicts a flare*):

$$FAR = 1 - PREC = \frac{FP}{TP + FP}$$

- *False Positive Rate (FPR)*: the fraction of positive predictions that are incorrect relative to the total number of negative events:

$$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}}$$

- *F1-score (F1)*: the harmonic mean of precision and recall, which balances false positives and false negatives:

$$\mathrm{F1} = 2 \cdot \frac{\mathrm{PREC} \cdot \mathrm{REC}}{\mathrm{PREC} + \mathrm{REC}} = \frac{2 \cdot \mathrm{TP}}{2 \cdot \mathrm{TP} + \mathrm{FP} + \mathrm{FN}}$$

- *Critical Success Index (CSI)*: the proportion of observed and/or predicted events that were correctly predicted:

$$\mathrm{CSI} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP} + \mathrm{FN}}$$

- *True Skill Statistic (TSS)*: *the difference between the probability of detection and the false positive rate*:

$$TSS = \frac{TP}{TP + FN} - \frac{FP}{FP + TN}$$

- *Heidke Skill Score (HSS)*: the fractional improvement of the forecast over random chance:

$$\mathrm{HSS} = \frac{2(\mathrm{TP} \cdot \mathrm{TN} - \mathrm{FN} \cdot \mathrm{FP})}{(\mathrm{TP} + \mathrm{FN})(\mathrm{FN} + \mathrm{TN}) + (\mathrm{TP} + \mathrm{FP})(\mathrm{FP} + \mathrm{TN})}$$

3.2. ROC Curve and Optimal Threshold

To evaluate the performance of the probabilistic forecasts across possible thresholds, we compute the Receiver Operating Characteristic (ROC) curve. The ROC curve plots the True Positive Rate (POD) against the False Positive Rate (FPR) as the decision threshold varies. The area under the ROC curve (AUC) quantifies the ability of the forecast to discriminate between flare and no-flare days; a perfect forecast has AUC = 1, while a no-skill forecast has AUC = 0.5. *Note that the TSS is the maximum vertical distance between the ROC curve and the diagonal line*.

3.3. Probabilistic Forecast Evaluation

In addition to evaluating binary classifications, we assess the quality of the raw probability forecasts. Two standard metrics are used:

- *Brier Score (BS)*: the mean squared error between forecast probabilities p_i and binary outcomes $y_i \in \{0, 1\}$:

$$\mathrm{BS} = \frac{1}{N}\sum_{i=1}^{N}(p_i - y_i)^2$$

  Lower Brier Scores indicate more accurate and better-calibrated probabilistic forecasts.

- *Reliability Diagram*: this plot compares the forecast probability to the observed frequency of flare occurrence. A well-calibrated forecast will fall along the diagonal; systematic deviations indicate overconfidence (below the diagonal) or underconfidence (above it).

3.4. Baseline Models and Context for Evaluation Metrics

To contextualize the performance of the SWPC forecasts, we compare them to several simple, interpretable models that require minimal computational cost and leverage only readily available solar indicators. These *zero-cost baselines* serve as reference points to determine whether operational forecasts provide added value beyond naive or climatological heuristics.

3.4.1. Persistence Model

The persistence model uses flare occurrence on the current day to predict flare occurrence on subsequent days. For each day D, the observed flare activity (M- or X-class) is used as the prediction for day $D + 1$, $D + 2$, and $D + 3$. While conceptually simple, this model can yield non-trivial skill during active solar periods when flare events are temporally clustered.

3.4.2. Climatology-Based and Statistical Models

We define three additional baselines that all rely on the same two input features:

- x_1: number of consecutive flare-free days prior to the forecast day;
- x_2: sunspot number on the forecast day.

  These features capture aspects of solar activity relevant to flare occurrence and are available in real time, making them suitable for operational use. The three models differ in how these inputs are used:

- *Empirical Climatology*: For each combination of flare-free days and sunspot number (x_1, x_2), we compute the empirical probability of a flare based on historical records. This conditional probability serves directly as the predicted probability of flare occurrence. The features are discretized into bins: $(0, 1, 2, \ldots, 20, > 20)$ for x_1, and $(0, 10, 20, \ldots, 200, > 200)$ for x_2. For each bin pair (b_1, b_2), we compute:

$$P_{\mathrm{clim}}\,(y = 1 | x_1 \in b_1, x_2 \in b_2) = \frac{N_{\mathrm{flare}}(b_1, b_2)}{N_{\mathrm{total}}(b_1, b_2)}, \quad (1)$$

where $N_{\text{flare}}(b_1, b_2)$ is the number of days with a flare in the bin pair, and $N_{\text{total}}(b_1, b_2)$ is the total number of days in that bin. This model reflects the empirically observed frequency of flares conditioned on recent activity and sunspot levels.

- *Naive Bayes Classifier (NB)*: Naive Bayes is a probabilistic classifier that assumes conditional independence between the input features given the class label $y \in \{0, 1\}$:

$$P(y = 1|x_1, x_2) = \frac{P(y = 1)P(x_1|y = 1)\,P(x_2|y = 1)}{\sum_{y' \in \{0,1\}} P(y')\,P(x_1|y')\,P(x_2|y')}. \tag{2}$$

The class priors $P(y = 1)$ and $P(y = 0)$ are computed from the data set. The feature likelihoods $P(x_1 \mid y)$ and $P(x_2 \mid y)$ can be estimated either using histograms or kernel density estimates. *Despite the strong assumption of feature independence (which is obviously not valid, given that sunspot number is not independent from number of consecutive flare-free days), Naive Bayes performs reasonably well. In this work, we use the Multinomial Naive Bayes implementation of scikit-learn* (Pedregosa et al., 2011).

- *Logistic Regression (LR)*: A linear probabilistic model in which the log-odds of flare occurrence is modeled as a function of the two input features. It provides a parametric alternative to the binned empirical climatology and can generalize better in regions of low data density.

$$\log\left(\frac{P(y = 1|x_1, x_2)}{1 - P(y = 1|x_1, x_2)}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2. \tag{3}$$

Equivalently, the probability is given by the sigmoid function:

$$P_{\text{LR}}(y = 1|x_1, x_2) = \frac{1}{1 + \exp\left(-\left(\beta_0 + \beta_1 x_1 + \beta_2 x_2\right)\right)}. \tag{4}$$

The model parameters $\beta_0, \beta_1, \beta_2$ are learned by maximizing the likelihood of the observed training data. Logistic regression captures the interaction between features in a parametric form and can generalize to unseen combinations better than the binned climatology model.

- *Baseline Average*: In addition to evaluating each model individually, we also compute a baseline average forecast, defined as the arithmetic mean of the predicted probabilities from four models: climatology, Naive Bayes, logistic regression, and persistence. Averaging forecasts is a common and effective ensemble strategy that helps reduce overfitting and model-specific biases. By averaging across these methods, the ensemble benefits from the diversity of its members and often achieves better calibration, lower variance, and improved robustness compared to any single model. This is particularly valuable in rare-event forecasting, where individual models may exhibit erratic behavior due to limited data.

3.5. Training Strategy

To ensure a fair and realistic comparison across all forecasting methods, we adopt a training strategy that strictly avoids the use of future information when generating predictions. Specifically, for the statistical and machine learning models (climatology, logistic regression, and Naive Bayes), we only use training data that would have been available up to the time each forecast was made. This simulates an operational scenario and prevents any leakage of future information into the training set.

We reserve the first 17 months of data (from 1996–2008 to 1997–2012) as a buffer period, used solely for model initialization and excluded from the evaluation. Beginning in January 1998, all models are retrained monthly: that is, each month's forecasts are produced using a model trained only on data from all previous months. This rolling retraining approach reflects how an operational system might continuously incorporate new data while remaining causally consistent.

While this retraining scheme is particularly important for machine learning models such as logistic regression and Naive Bayes to avoid overfitting, it is, in principle, less critical for the climatology model. Climatology is typically defined as a static method based on long-term averages. However, for methodological consistency, we

apply the same monthly retraining procedure to the climatology model as well. Persistence, on the contrary, does not require any training.

3.6. Thresholding for Binary Metrics

The SWPC forecast and all of the baseline models (except for persistence) produce probabilistic outputs, that is, $P(y = 1 \mid x_1, x_2)$. To evaluate classification performance using binary metrics, we must apply a decision threshold θ:

$$\hat{y} = \begin{cases} 1 & \text{if } P(y = 1 | x_1, x_2) \geq \theta, \\ 0 & \text{otherwise.} \end{cases} \tag{5}$$

In our analysis, we explore two scenarios: (a) The standard choice for a well-calibrated model is to set $\theta = 0.5$. *This is what a typical user would choose, particularly given that the NOAA SWPC website does not provide any alternative information. (b) Alternatively, one can choose the threshold that maximizes specific metrics. Here, we choose to maximize True Skill Statistic (TSS).*

3.7. Interpreting Metrics in Imbalanced Contexts

The extreme imbalance between flare and no-flare days poses challenges for evaluating forecast skill. *For example, in our data set of 9,828 days, M-class flares occurred only on* 20.6% *of days and X-class flares only on* 2.6%. *As a result, a model that always predicts no flare achieves* 79.4% *accuracy for M-class and* 97.4% *for X-class*. Despite appearing high, this reflects a trivial solution and provides no actionable forecasting value, *hence the need to look at other metrics, such as Probability of Detection (POD, or Recall), Critical Success Index (CSI), and Heidke Skill Score (HSS) as key metrics in this study. The Brier score, which is often used to assess probabilistic forecast suffers the same shortfall: a model that always predicts no flare has a Brier score of 0.026 for X-class flares in our data set (Brier = 0 is a perfect prediction).*

High recall often comes at the cost of low precision, especially for rare events like X-class flares. We therefore also examine the False Alarm Ratio (FAR) and Precision to capture this tradeoff. While a high recall may be desirable for operational purposes (e.g., safety), a high false alarm rate can erode trust in forecasts. *It is worth noting that in a highly imbalanced, mostly TN, data set, FPs can be "hidden" by the huge number of TNs. In this scenario, TSS is compromised since FPR tends to 0 even though FPs are very high thus giving a misleading sense of high TSS. Many flare prediction papers fall prey to this and report relatively high TSS values (e.g.,* >0.8) *from their model while failing to note that the FAR is extremely high, therefore making their model unsuitable for operational forecasting*.

Together, these baseline models and evaluation strategies provide a rigorous framework to assess the SWPC operational forecasts *(or any other flare forecasting method with sufficient data for statistical analysis) using simple, interpretable baselines based on easily accessible solar activity features*.

4. Results

4.1. Reliability of SWPC Probabilistic Forecasts

We begin our analysis by evaluating the reliability of the SWPC forecasts using reliability diagrams, which compare the forecast probabilities with the observed frequency of events. A forecast is considered *reliable* if, over all days when a certain probability p is forecast, a flare occurs approximately p% of the time.

Unlike many probabilistic forecast systems that issue probabilities with fine granularity, the SWPC forecasts are limited to a discrete set of probability values. Specifically, each lead time and flare class uses a fixed list of possible probabilities (percentage):

- *M-class forecasts (24hr, 48hr, 72hr): {1, 5, 10, 15, …, 95} (20 values)*
- *X-class 24hr: {1, 2, 5, 10, …, 75} (except for 65, 70–15 values)*
- *X-class 48hr: {1, 2, 5, 10, …, 75} (except for 65–16 values)*
- *X-class 72hr: {1, 2, 5, 10, …, 75} (except for 70–16 values)*

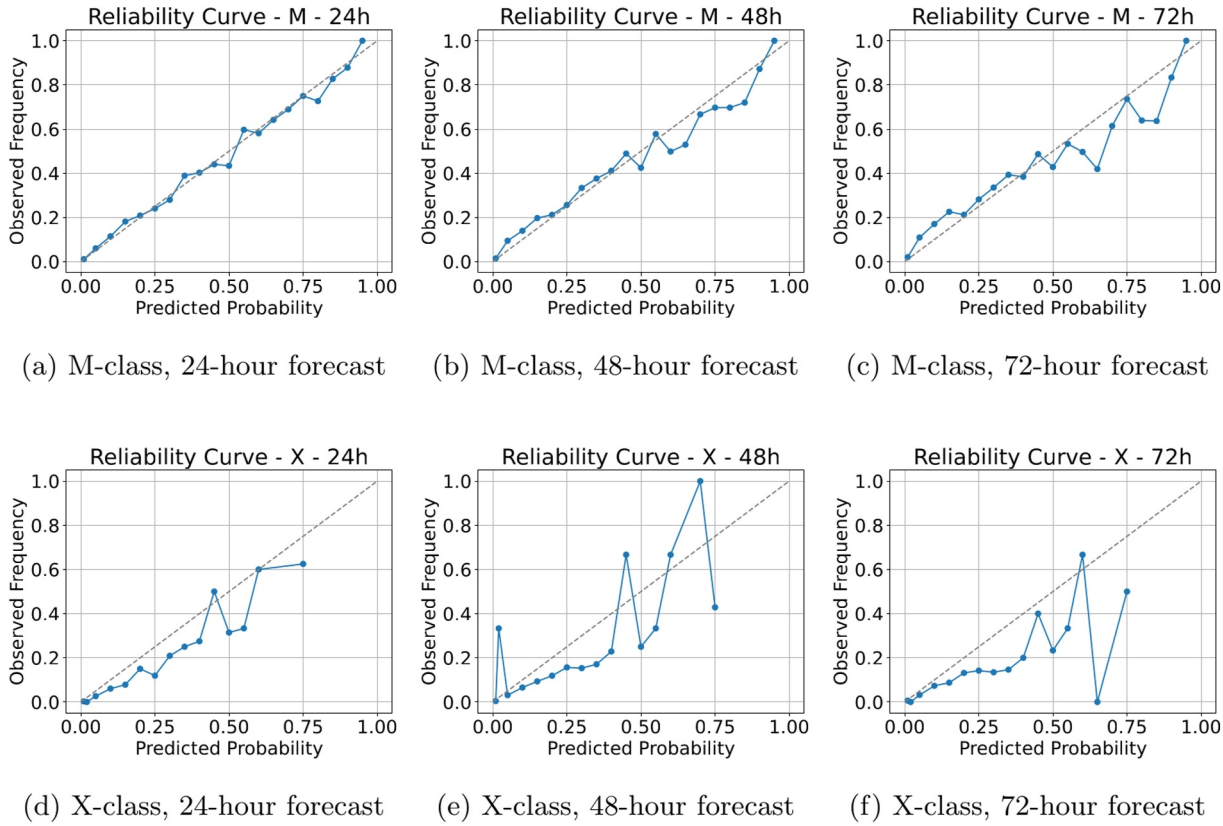

Figure 4. Reliability diagrams for M-class and X-class flare forecasts at 24-hr, 48-hr, and 72-hr lead times. The diagonal dashed line represents perfect reliability.

In passing, we notice that on the day 2007-09-04 the issued probabilities were 71/71/71 for M-class, which we considered a typo and changed to 70/70/70 (being the only one instance of 71% probability in the whole database).

As a result, we plot the observed relative frequencies directly against these discrete forecast values, rather than grouping into broader bins. Each point on the diagram corresponds to one of the forecast probability values, and its vertical position reflects the empirical probability of flare occurrence given that forecast value. *Perfect reliability corresponds to the diagonal $y = x$ line, where the predicted probability equals the observed frequency.*

M-class flare forecasts: *The top row of Figure 4 shows the reliability diagrams for M-class forecasts at 24-hr, 48-hr, and 72-hr lead times. At 24-hr lead time, the forecasts are generally well-calibrated over the whole probability range. At longer lead times, forecast probabilities become less reliable, with a tendency toward overconfidence. For example, the 72-hr forecast at the predicted probability 80% has an observed frequency of only 60%.*

X-class flare forecasts: The reliability diagrams for the X-class forecasts are shown in the bottom row of Figure 4. Due to the rarity of X-class flares, the reliability curves exhibit greater variance and less smoothness, especially at higher probability bins where data are sparse. At all lead times, forecasts tend to be overconfident: the predicted probability exceeds the observed frequency of events, especially for probabilities above 20%. This is consistent with a high FAR and is expected in rare-event forecasting when attempting to maximize recall. *The 48-hr forecast (middle column) shows an inconsistent behavior with the reliability curve crossing the diagonal line several times.*

The discrete nature of the forecast probabilities limits the resolution of the reliability curves, but also simplifies interpretation. Overall, the reliability analysis reveals that M-class forecasts are reasonably well-calibrated at short lead times, while X-class forecasts tend to be overconfident and degrade more quickly with increasing lead time. These patterns should be considered when interpreting the deterministic performance metrics presented in the next sections.

Table 2
Comparison of Flare Prediction Models Across Metrics M Class, 24 hr Ahead (Clim. = Climatology; Pers. = Persistence; NB = Naive Bayes; LR = Logistic Regression; BA = Baseline Average)

Model	Accuracy	Precision	Recall	F1	Brier	AUC	CSI	POD	FAR	TSS	HSS
SWPC	**0.84**	0.62	0.53	0.57	**0.11**	**0.87**	0.40	0.53	0.38	0.44	**0.47**
Clim.	0.80	0.53	0.34	0.42	0.13	0.77	0.26	0.34	0.47	0.26	0.30
Pers.	0.82	0.57	0.57	0.57	0.18	0.73	0.40	0.57	0.43	0.46	0.46
NB	0.63	0.35	**0.93**	0.51	0.36	0.79	0.34	**0.93**	0.65	**0.48**	0.30
LR	0.81	**0.64**	0.15	0.24	0.13	0.83	0.13	0.15	**0.36**	0.12	0.17
BA	0.82	0.57	0.59	**0.58**	0.12	0.86	**0.41**	0.59	0.43	0.47	**0.47**

Note. Bold text indicates the best model for that metric.

4.2. Standard Threshold $\theta = 0.5$

Here, we evaluate flare prediction performance across multiple models and forecast lead times (24-, 48-, and 72-hr) for both M-class and X-class events, using a standard probability threshold of 0.5. Tables 2–7 summarize the results. Overall, the NOAA SWPC forecast is consistently outperformed by baseline models across nearly all skill scores—most notably for metrics less sensitive to class imbalance, such as the F1, HSS, and CSI.

4.2.1. Performance on M-Class Flare Forecasts

Across all lead times (24hr, 48hr, and 72hr), the SWPC model achieves the highest accuracy and lowest Brier scores. However, these metrics can be misleading in highly imbalanced data sets like solar flare occurrence, where non-events dominate. High accuracy may reflect a tendency to predict "no flare" by default, and a low Brier score can indicate cautious probability estimates rather than true predictive skill.

When focusing on event-focused metrics—such as F1 score, Critical Success Index (CSI), and Heidke Skill Score (HSS)—the Baseline Average (BA) model consistently outperforms SWPC at all lead times. At 24 hr, BA achieves the best F1 score (0.58), CSI (0.41), and matches SWPC in HSS (0.47), while offering higher recall (0.59) with comparable precision. Similar trends persist at 48 and 72 hr, where BA maintains stronger overall balance between detecting flares and avoiding false alarms.

Remarkably, even the simple persistence model—using no training and based solely on the previous day's flare activity—performs on par with, or only marginally below, the SWPC forecast. At all lead times, persistence achieves similar or better scores than SWPC in recall, F1, CSI, and HSS.

4.2.2. X-Class Flares

The performance gap is even more pronounced for the rare but operationally critical X-class flares. The SWPC forecast exhibits low recall ($\leq$0.08) and corresponding CSI values ($\leq$0.07) at all forecast horizons, indicating poor detection rates.

Table 3
Comparison of Flare Prediction Models Across Metrics M Class, 48 hr Ahead (Clim. = Climatology; Pers. = Persistence; NB = Naive Bayes; LR = Logistic Regression; BA = Baseline Average)

Model	Accuracy	Precision	Recall	F1	Brier	AUC	CSI	POD	FAR	TSS	HSS
SWPC	**0.82**	0.57	0.46	0.51	**0.12**	**0.85**	0.34	0.46	0.43	0.37	0.40
Clim.	0.80	0.52	0.24	0.33	0.14	0.75	0.20	0.24	0.48	0.19	0.23
Pers.	0.81	0.53	0.53	0.53	0.19	0.70	0.36	0.53	0.47	0.41	0.41
NB	0.62	0.34	**0.91**	0.50	0.37	0.78	0.33	**0.91**	0.66	**0.46**	0.29
LR	0.81	**0.64**	0.13	0.22	0.13	0.82	0.12	0.13	**0.36**	0.11	0.16
BA	0.81	0.53	0.55	**0.54**	0.13	0.84	**0.37**	0.55	0.47	0.42	**0.42**

Note. Bold text indicates the best model for that metric.

Table 4
Comparison of Flare Prediction Models Across Metrics M Class, 72 hr Ahead (Clim. = Climatology; Pers. = Persistence; NB = Naive Bayes; LR = Logistic Regression; BA = Baseline Average)

Model	Accuracy	Precision	Recall	F1	Brier	AUC	CSI	POD	FAR	TSS	HSS
SWPC	**0.81**	0.55	0.43	0.48	**0.13**	**0.83**	0.32	0.43	0.45	0.34	0.37
Clim.	0.79	0.49	0.21	0.29	0.14	0.74	0.17	0.21	0.51	0.15	0.19
Pers.	0.79	0.49	0.49	0.49	0.21	0.68	0.33	0.49	0.51	0.36	0.36
NB	0.62	0.34	**0.90**	0.49	0.38	0.77	0.33	**0.90**	0.66	**0.45**	0.28
LR	**0.81**	**0.67**	0.11	0.19	**0.13**	0.81	0.11	0.11	**0.33**	0.10	0.14
BA	0.80	0.50	0.51	**0.51**	0.14	0.82	**0.34**	0.51	0.50	0.38	**0.38**

Note. Bold text indicates the best model for that metric.

Across all lead times (24hr, 48hr, 72hr), the SWPC model achieves the highest accuracy (0.97) and the lowest Brier scores (0.02), and maintains the best AUC values (up to 0.87). However, as with M-class flares, these metrics can be misleading in highly imbalanced data sets. The SWPC model exhibits very low recall (as low as 0.05–0.08) and F1 scores (0.08–0.13), meaning it fails to capture most X-class flare events. The result is a model that rarely predicts flares and is therefore of limited use for event detection.

In contrast, the persistence model achieves significantly better balance in key operational metrics. It consistently outperforms SWPC in recall (0.19–0.21), F1 (up to 0.21), Critical Success Index (CSI), and both TSS and HSS, despite having no underlying data model. For example, at 24hr lead time, persistence achieves the highest CSI (0.12) and HSS (0.19), highlighting its effectiveness as a zero-cost benchmark. The Baseline Average (BA), constructed here by averaging the predictions of climatology, persistence, and Naive Bayes, consistently performs more robustly than SWPC (*logistic regression is excluded in this case for its low performance*). While it does not lead in any individual metric, it achieves a more balanced compromise between precision and recall, yielding higher F1 scores and CSI values than climatology or SWPC at all lead times. Notably, at 72hr, BA matches persistence in HSS (0.16), with a similar CSI (0.10) and recall (0.17), and outperforms SWPC in F1, CSI, and HSS. These results reinforce the idea that even simple ensemble strategies can outperform or rival complex, opaque forecasting systems in rare event prediction.

Overall, the SWPC forecast for X-class flares shows poor event detection skill and is routinely outperformed by persistence and ensemble baselines. This underscores the need for operational models to prioritize recall and discrimination when dealing with rare but high-impact events.

4.2.3. Summary of Forecast Comparisons

The comprehensive validation of flare forecasts across M-class and X-class events, and for lead times of 24, 48, and 72 hr, reveals consistent patterns in model performance. For M-class flares, the SWPC forecast achieves the highest accuracy and Brier scores, but these metrics are inflated by class imbalance and do not reflect true event detection skill. In contrast, the Baseline Average (BA) model—a simple ensemble of zero-cost models—consistently achieves superior scores in event-sensitive metrics such as F1, CSI, TSS, and HSS, indicating better balance between

Table 5
Comparison of Flare Prediction Models Across Metrics X Class, 24 hr Ahead (Clim. = Climatology; Pers. = Persistence; NB = Naive Bayes; BA = Baseline Average)

Model	Accuracy	Precision	Recall	F1	Brier	AUC	CSI	POD	FAR	TSS	HSS
SWPC	**0.97**	**0.39**	0.08	0.13	**0.02**	**0.87**	0.07	0.08	**0.61**	0.08	0.12
Clim.	**0.97**	0.10	0.03	0.04	0.03	0.60	0.02	0.03	0.90	0.02	0.03
Pers.	0.96	0.21	0.21	**0.21**	0.04	0.60	**0.12**	0.21	0.79	0.19	**0.19**
NB	0.57	0.05	**0.84**	0.09	0.43	0.74	0.05	**0.84**	0.95	**0.40**	0.04
BA	0.96	0.19	0.18	0.18	0.07	0.80	0.10	0.18	0.81	0.16	0.16

Note. Bold text indicates the best model for that metric.

Table 6
Comparison of Flare Prediction Models Across Metrics X Class, 48 hr Ahead (Clim. = Climatology; Pers. = Persistence; NB = Naive Bayes; BA = Baseline Average)

Model	Accuracy	Precision	Recall	F1	Brier	AUC	CSI	POD	FAR	TSS	HSS
SWPC	**0.97**	**0.33**	0.06	0.10	**0.02**	**0.84**	0.05	0.06	**0.67**	0.06	0.09
Clim.	**0.97**	0.06	0.01	0.02	0.03	0.57	0.01	0.01	0.94	0.01	0.01
Pers.	0.96	0.20	0.20	**0.20**	0.04	0.59	**0.11**	0.20	0.80	0.18	**0.18**
NB	0.56	0.05	**0.82**	0.09	0.44	0.73	0.05	**0.82**	0.95	**0.37**	0.04
BA	0.96	0.19	0.18	0.19	0.07	0.77	0.10	0.18	0.81	0.16	0.16

Note. Bold text indicates the best model for that metric.

sensitivity and specificity. Notably, even the persistence model, which requires no training and uses no flare history beyond the previous day, performs on par with or slightly below SWPC in most skill scores.

For X-class flares, the limitations of the SWPC forecast become even more pronounced. While it maintains the highest accuracy and lowest Brier scores, its recall is very low and it fails to detect most flare events, resulting in low F1 and CSI scores. Persistence again outperforms SWPC in almost all event-relevant metrics, while Naive Bayes shows high recall but unacceptably high false alarm rates. The Baseline Average model offers a well-balanced alternative, outperforming SWPC in key scores like F1, CSI, and HSS at all horizons. Logistic regression was excluded from the X-class comparison due to poor calibration and negligible recall.

Overall, the findings suggest that SWPC forecasts do not significantly outperform zero-cost or simple statistical baselines, particularly in rare-event detection.

It is important to stress that comparisons with zero-cost baseline models are not meant to suggest these models as viable forecasting alternatives on their own, but rather that they should be considered as additional tools for human forecasters when issuing 1–3 days flare forecasts. Ideally, an official forecast should be developed as a multi-model ensemble with human-in-the-loop interpretation that takes into account the skill of experienced space weather forecasters developed over many years of observing solar behavior.

4.3. Optimal Probability Threshold

In this section, we determine the probability threshold θ that optimizes the TSS metric for each model independently. The optimal values of θ are summarized in Table 8. We recall that this value sets the threshold for positive/negatives for the binary classification metrics. Typically, a well-calibrated model would have an optimal threshold close to 0.5.

The optimal probability thresholds presented in Table 8 highlight substantial calibration issues with the SWPC forecasts, particularly for X-class flares, where thresholds as low as 0.05 are required to produce any positive forecasts. These extremely low thresholds indicate that SWPC X-class flare forecasts are systematically under-confident, necessitating aggressive post-processing to extract meaningful predictions. In contrast, the baseline average model exhibits more moderate and consistent thresholds, suggesting better calibration (*the threshold for persistence is either zero or one by construction*).

Table 7
Comparison of Flare Prediction Models Across Metrics X Class, 72 hr Ahead (Clim. = Climatology; Pers. = Persistence; NB = Naive Bayes; BA = Baseline Average)

Model	Accuracy	Precision	Recall	F1	Brier	AUC	CSI	POD	FAR	TSS	HSS
SWPC	**0.97**	**0.29**	0.05	0.08	**0.02**	**0.81**	0.04	0.05	**0.71**	0.04	0.07
Clim.	**0.97**	0.06	0.01	0.02	0.03	0.53	0.01	0.01	0.94	0.01	0.01
Pers.	0.96	0.19	0.19	**0.19**	0.04	0.58	**0.10**	0.19	0.81	0.16	**0.16**
NB	0.55	0.04	**0.79**	0.08	0.44	0.72	0.04	**0.79**	0.96	**0.34**	0.04
BA	0.96	0.18	0.17	0.18	0.07	0.73	**0.10**	0.17	0.82	0.15	**0.16**

Note. Bold text indicates the best model for that metric.

Table 8
Optimal Values of Probability Threshold θ, Defined as the Value That Maximizes TSS

Model	M 24 hr	M 48 hr	M 72 hr	X 24 hr	X 48 hr	X 72 hr
SWPC	0.20	0.20	0.15	0.05	0.05	0.05
Climatology	0.15	0.14	0.16	0.03	0.03	0.03
Persistence	1.00	1.00	1.00	1.00	1.00	1.00
Naive Bayes	1.00	1.00	1.00	1.00	1.00	1.00
Logistic Regression	0.18	0.18	0.17	0.03	0.02	0.03
Baseline Average	0.36	0.36	0.34	0.26	0.26	0.26

Applying optimized probability thresholds for binary classification improves the event detection capabilities of all models, leading to a more meaningful evaluation of forecast skill under imbalanced conditions. The results for M-class flares at 24, 48, and 72 hr (Tables 9–11) reveal that the SWPC model improves its recall substantially (e.g., 0.86 at 24hr) while maintaining a low Brier score and high AUC, suggesting good calibration and discrimination. However, when judged by event-focused metrics such as F1, CSI, TSS, and HSS, the performance gap between SWPC and baseline models narrows considerably—and in some cases, disappears.

Performance comparisons across all forecast lead times and flare classes (Tables 9–14) consistently show that baseline models—particularly climatology and persistence—either match or outperform the SWPC model on most metrics. While SWPC tends to achieve high recall values (up to 0.93 for X-class flares at 24 hr; Table 12), this comes at the cost of extremely high false alarm rates (FARs), often exceeding 90%. These inflated FARs greatly reduce the practical utility of the SWPC forecasts.

Persistence, while simplistic, achieves the highest accuracy and precision in many settings (e.g., X-class, all lead times; Tables 12–14), with much lower FARs than SWPC, despite its lower recall.

The poor calibration and elevated FARs associated with the SWPC model suggest that the forecasts, in their current form, are not well-suited for operational decision-making without significant recalibration or supplementation with baseline or statistical methods.

Overall, the use of optimized thresholds reveals that SWPC's flare forecast is closely matched or exceeded by simple statistical models and the baseline average. The persistence model, in particular, continues to provide competitive skill with no learning or parameter tuning, emphasizing the need for operational models to demonstrate clear added value over trivial heuristics.

4.4. Best Trade-Off Across Metrics

When considering all metrics jointly—particularly those most relevant for operational performance such as F1 score, Critical Success Index (CSI), True Skill Statistic (TSS), and Heidke Skill Score (HSS)— SWPC forecasts are not consistently outperforming the **Baseline Average (BA)** for M-class flares and **Persistence** for X-class flares.

Table 9
Comparison of Flare Prediction Models Across Metrics M Class, 24 hr Ahead (Optimized Threshold) (Clim. = Climatology; Pers. = Persistence; NB = Naive Bayes; BA = Baseline Average)

Model	Accuracy	Precision	Recall	F1	Brier	AUC	CSI	POD	FAR	TSS	HSS
SWPC	0.75	0.44	0.86	**0.58**	**0.11**	**0.87**	**0.41**	0.86	0.56	**0.58**	0.43
Clim.	0.72	0.41	0.80	0.54	0.13	0.77	0.37	0.80	0.59	0.50	0.37
Pers.	**0.82**	**0.57**	0.57	0.57	0.18	0.73	0.40	0.57	**0.43**	0.46	**0.46**
NB	0.67	0.37	**0.89**	0.52	0.36	0.79	0.35	**0.89**	0.63	0.50	0.33
LR	0.72	0.42	0.85	0.56	0.13	0.83	0.39	0.85	0.58	0.54	0.39
BA	0.76	0.46	0.81	**0.58**	0.12	0.86	**0.41**	0.81	0.54	0.56	0.43

Note. Bold text indicates the best model for that metric.

Table 10
Comparison of Flare Prediction Models Across Metrics M Class, 48 hr Ahead (Optimized Threshold) (Clim. = Climatology; Pers. = Persistence; NB = Naive Bayes; BA = Baseline Average)

Model	Accuracy	Precision	Recall	F1	Brier	AUC	CSI	POD	FAR	TSS	HSS
SWPC	0.74	0.43	0.83	**0.56**	**0.12**	**0.85**	**0.39**	0.83	0.57	**0.54**	0.40
Clim.	0.69	0.38	0.80	0.51	0.14	0.75	0.35	0.80	0.62	0.46	0.33
Pers.	**0.81**	**0.53**	0.53	0.53	0.19	0.70	0.36	0.53	**0.47**	0.41	**0.41**
NB	0.66	0.36	**0.87**	0.51	0.37	0.78	0.34	**0.87**	0.64	0.47	0.31
LR	0.71	0.40	0.83	0.54	0.13	0.82	0.37	0.83	0.60	0.51	0.37
BA	0.75	0.44	0.78	**0.56**	0.13	0.84	**0.39**	0.78	0.56	0.52	0.40

Note. Bold text indicates the best model for that metric.

For M-class flare prediction, the Baseline Average consistently achieves top or near-top performance across F1, CSI, and HSS for all forecast horizons (24hr, 48hr, and 72hr). It combines high recall with moderate false alarm rates and outperforms or matches the SWPC forecast in all threshold-dependent metrics. The ensemble nature of BA allows it to capture the strengths of its component models (climatology, Naive Bayes, and logistic regression), leading to more balanced and reliable classification performance.

For X-class flares, Persistence emerges as the most reliable and balanced model. While the SWPC forecast achieves extremely high recall, it does so at the expense of precision, resulting in very high false alarm rates and low F1 and CSI scores. In contrast, Persistence offers moderate recall (0.19–0.21), substantially higher precision (0.19–0.21), and the best F1, CSI, and HSS values across all lead times. It achieves this without any learning or model complexity, highlighting the importance of evaluating against strong baseline methods.

4.5. Storm After the Calm: Forecasting the First Event After Extended Quiet

A particularly challenging scenario for flare forecasting is the detection of a sudden, isolated X-class flare following a prolonged period of solar quiet. This situation, which we term the *storm after the calm*, is especially relevant for human spaceflight, where early warning is critical to ensure astronauts can seek shelter in time. *To assess this, we focus on a particularly demanding scenario: days for which the number of prior flare-free days is larger than 30, meaning the Sun has been quiet for over a month. In this regime, we evaluate whether the model can detect the first sudden X-class flare after a prolonged period of inactivity*. The subset contains 66 positive cases (X-flares) and 6,992 negatives, making it both operationally relevant and extremely imbalanced.

Using the SWPC model's optimal threshold for X-class flares at 24 hr (5%), the confusion matrix is shown in Figure 5 (TP = 56, FN = 10, TN = 5,735, FP = 1,257).

The results are concerning. Ten of the X-class flares were missed—15% or 1 in every 7 dangerous events. Each of these false negatives represents a potentially catastrophic failure. In human spaceflight, one missed X-flare can be lethal. A model that fails to anticipate such events, even occasionally, cannot be trusted in high-stakes operational contexts.

Table 11
Comparison of Flare Prediction Models Across Metrics M Class, 72 hr Ahead (Optimized Threshold) (Clim. = Climatology; Pers. = Persistence; NB = Naive Bayes; BA = Baseline Average)

Model	Accuracy	Precision	Recall	F1	Brier	AUC	CSI	POD	FAR	TSS	HSS
SWPC	0.71	0.40	0.84	**0.54**	**0.13**	**0.83**	**0.37**	0.84	0.60	**0.51**	**0.36**
Clim.	0.69	0.37	0.76	0.50	0.14	0.74	0.33	0.76	0.63	0.43	0.31
Pers.	**0.79**	**0.49**	0.49	0.49	0.21	0.68	0.33	0.49	**0.51**	0.36	**0.36**
NB	0.65	0.36	**0.86**	0.50	0.38	0.77	0.34	**0.86**	0.64	0.46	0.30
LR	0.69	0.39	0.84	0.53	**0.13**	0.81	0.36	0.84	0.61	0.50	0.35
BA	0.71	0.40	0.79	0.53	0.14	0.82	0.36	0.79	0.60	0.48	0.35

Note. Bold text indicates the best model for that metric.

Table 12
Comparison of Flare Prediction Models Across Metrics X Class, 24 hr Ahead (Optimized Threshold) (Clim. = Climatology; Pers. = Persistence; NB = Naive Bayes; BA = Baseline Average)

Model	Accuracy	Precision	Recall	F1	Brier	AUC	CSI	POD	FAR	TSS	HSS
SWPC	0.69	0.07	**0.93**	0.14	**0.02**	**0.87**	0.07	**0.93**	0.93	**0.61**	0.09
Clim.	0.88	0.09	0.38	0.14	0.03	0.60	0.08	0.38	0.91	0.27	0.10
Pers.	**0.96**	**0.21**	0.21	**0.21**	0.04	0.60	**0.12**	0.21	**0.79**	0.19	**0.19**
NB	0.62	0.05	0.80	0.10	0.43	0.74	0.05	0.80	0.95	0.42	0.05
LR	0.72	0.06	0.69	0.11	**0.02**	0.77	0.06	0.69	0.94	0.42	0.07
BA	0.61	0.05	0.84	0.10	0.05	0.80	0.05	0.84	0.95	0.45	0.06

Note. Bold text indicates the best model for that metric.

Moreover, the SWPC forecast produces a large number of false alarms: 1,257 quiet days incorrectly predicted as flare-active out of 1,313 predicted positive (*95% false alarm ratio), undermining its credibility and potentially leading to mission planners choosing to ignore the forecast altogether*.

4.6. All-Clear Events: Forecasting a Return to Quiet After X-Class Flares

Complementary to the "storm after the calm" scenario, we evaluate the ability of the SWPC forecast to identify *all-clear* periods following an X-class flare. Specifically, we define an all-clear event as a day with no X-class flares occurring **+1, +2, or +3 days** after an X-class flare. For each of these days, we assess the model's 24-hr forecast issued the day before.

This is particularly crucial for human spaceflight (Ji et al., 2020; Sadykov et al., 2021). Planning for Extravehicular Activities (EVAs) or sorties on the surface of the Moon or Mars, during which astronauts have a greatly enhanced risk of radiation exposure due to the lack of spacecraft or habitat shielding, relies on identifying "flare-free" periods days in advance. In these activities, a false negative, or missed, X-class flare could result in lethal radiation exposure to the astronauts on the EVA. Therefore, the ability to accurately and reliably predict extended quiet periods, and to forecast the first sign of dangerous activity, is a defining test of a model's usefulness.

We construct a subset of forecast instances by selecting all days that are +1, +2, or +3 after a confirmed X-class flare, yielding a data set focused specifically on the period immediately following high activity. For each instance, we apply the SWPC 24-hr forecast with its optimal threshold for X-class events (5%) to determine whether the model predicts renewed flare activity.

The confusion matrix for all-clear periods is shown in Figure 6 *(TP = 150, FN = 1, TN = 38, FP = 576)*.

The forecast successfully identifies nearly all flare events (recall = 0.99), but it performs very poorly at identifying true all-clear days (precision = 0.21, FAR = 0.79). Of the 614 non-flaring days, only 38 are correctly predicted as quiet. The overwhelming number of false positives—576 out of 614—indicates a strong bias toward over-warning, severely limiting the forecast's value as a tool for operational confidence.

Table 13
Comparison of Flare Prediction Models Across Metrics X Class, 48 hr Ahead (Optimized Threshold) (Clim. = Climatology; Pers. = Persistence; NB = Naive Bayes; BA = Baseline Average)

Model	Accuracy	Precision	Recall	F1	Brier	AUC	CSI	POD	FAR	TSS	HSS
SWPC	0.70	0.07	**0.88**	0.13	**0.02**	**0.84**	0.07	**0.88**	0.93	**0.58**	0.09
Clim.	0.80	0.05	0.41	0.10	0.03	0.57	0.05	0.41	0.95	0.22	0.05
Pers.	**0.96**	**0.20**	0.20	**0.20**	0.04	0.59	**0.11**	0.20	**0.80**	0.18	**0.18**
NB	0.64	0.05	0.75	0.10	0.44	0.73	0.05	0.75	0.95	0.38	0.05
LR	0.57	0.05	0.84	0.09	**0.02**	0.76	0.05	0.84	0.95	0.40	0.04
BA	0.62	0.05	0.80	0.10	0.05	0.78	0.05	0.80	0.95	0.41	0.05

Note. Bold text indicates the best model for that metric.

Table 14
Comparison of Flare Prediction Models Across Metrics X Class, 72 hr Ahead (Optimized Threshold) (Clim. = Climatology; Pers. = Persistence; NB = Naive Bayes; BA = Baseline Average)

Model	Accuracy	Precision	Recall	F1	Brier	AUC	CSI	POD	FAR	TSS	HSS
SWPC	0.71	0.07	**0.83**	0.13	**0.02**	**0.81**	0.07	**0.83**	0.93	**0.53**	0.08
Clim.	0.80	0.04	0.32	0.08	0.03	0.53	0.04	0.32	0.96	0.13	0.03
Pers.	**0.96**	**0.19**	0.19	**0.19**	0.04	0.58	**0.10**	0.19	**0.81**	0.16	**0.16**
NB	0.63	0.05	0.73	0.09	0.44	0.72	0.05	0.73	0.95	0.36	0.05
LR	0.67	0.05	0.72	0.10	**0.02**	0.75	0.05	0.72	0.95	0.38	0.05
BA	0.62	0.05	0.75	0.09	0.05	0.75	0.05	0.75	0.95	0.36	0.05

Note. Bold text indicates the best model for that metric.

In summary, although the SWPC forecast performs moderately well in aggregate statistics, its behavior in both all-clear and storm-after-the-calm scenarios reveals a fundamental limitation: it is not yet reliable enough *to be used in actionable decisions regarding human spaceflight safety. NASA is currently funding a center of excellence for developing new all-clear radiation event forecasting models (Zhao, 2023). It will be imperative to compare the model(s) developed by this center objectively using the methods and metric described here (as well as others) to demonstrate improved forecast efficacy for the critical all-clear condition.*

5. Conclusions

This study presents a comprehensive verification of the NOAA/SWPC operational forecasts for M-class and X-class solar flares over the period 1998–2024. We find that SWPC forecasts perform comparably to, or are outperformed by, simple statistical baselines—including persistence, climatology, and lightweight machine learning classifiers such as logistic regression and Naive Bayes—across a wide range of skill metrics. In many cases, particularly for X-class flares, SWPC forecasts demonstrate poor calibration, low precision, and high false alarm rates, *emphasizing the need to move beyond a single model or method for the issuance of official solar flare forecasts*.

The findings shown here are concerning in high-risk/high-impact operational contexts. Of particular concern is the model's performance in "all-clear" scenarios, where its inability to reliably detect the onset of dangerous X-class flares after prolonged quiet periods could lead to *life-threatening outcomes for astronauts executing EVAs or in unsheltered surface exploration missions on the Moon or Mars*.

We suggest that in addition to the current reliance solely on the McIntosh classification-based method, SWPC and other operational space weather forecasting offices should adopt a multi-model approach, using some or all of the comparison models shown here in an ensemble approach that produces a more reliably skilled flare forecast. In particular, operational forecasting offices should work toward transitioning both the simple baselines shown here as well as more advanced state-of-the-art ML flare prediction models once such models have been demonstrated to produce verifiably accurate and reliable predictions of solar flares.

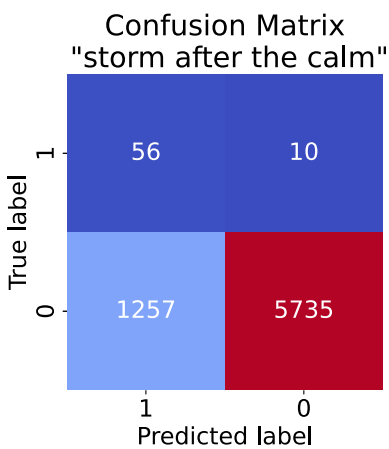

Figure 5. Confusion matrix for "storm after the calm" scenario (X flares).

In reference to this latter condition, any solar flare forecasting or all-clear prediction models developed in a research setting should perform the type of basic forecast verification study shown here—using the same baseline models and metrics for comparison—before being claimed as an advance over current methods. Many recent models developed for solar flare prediction have been published with, in some cases, TSS > 0.9 *and yet without reporting of the FAR, FPR, or other relevant "negative metrics" over a sufficiently long period to provide a credible evaluation of skill. While we do not wish to single out particular models for criticism, it is worth emphasizing that in highly imbalanced event classes like solar flares*, TN is often ≫ TP or FP *and thus the FPR is effectively zero, the TSS is artificially high (essentially just Recall), and extremely high FARs are effectively ignored. Indeed, specific model*

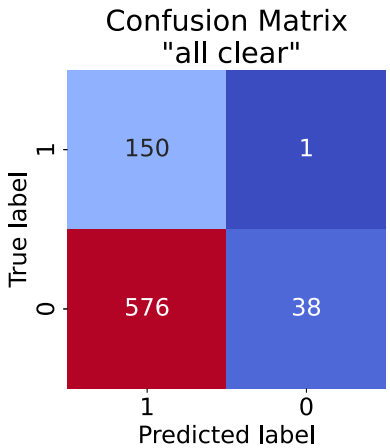

Figure 6. Confusion matrix for "all clear" scenario (X flares).

architectures have been developed to address the "disguised FPR" problem in solar flare prediction (Deshmukh et al., 2022).

While meta-studies like Leka et al. (2019) are useful for comparing the skills of a fixed set of models for a set period of solar activity, it should be the responsibility of the model developers themselves, or perhaps a neutral third-party such as NASA's Community Coordinated Modeling Center (CCMC), to run extensive and standardized forecast verification studies like the one demonstrated here prior to publishing a new prediction model and claiming superior predictive performance.

We emphasize that forecast verification and model validation are separate and very different activities. Model validation is usually a "nowcast verification" where input to a model at a given time produces model state output that is compared to observations at that same time, that is, with no forecast of the future system state. Model validation exercises often include extensive pre-processing of input data, pre-selection of favorable conditions and/or elimination of unfavorable conditions, and model tuning between runs to optimize performance on a given data set, making such exercises essentially irrelevant to actual operational forecasting use. It is possible (and apparently not uncommon) for a space weather model to validate well and yet demonstrate low forecasting skill.

In closing, we encourage all operational forecasting offices to (a) transition both baselines and state-of-the-art machine learning models, which have demonstrated robust performance and can be implemented at low computational cost, to operations as soon as practicable to supplement existing forecasting methods, and (b) routinely perform and publish rigorous operational forecast verifications using open, reproducible methods and transparent metrics. Only by doing so can we build trust among the space weather end-user community and ensure that operational space weather products meet the demands of critical infrastructure applications.

Appendix A: On the Use of Generative AI in This Work

Both authors are enthusiastic supporters of the responsible use of artificial intelligence in research and publishing. We recognize both the transformative potential and the ethical challenges that generative AI tools pose to academic work. As scientists and as members of the editorial community at AGU, we believe that thoughtful experimentation with these technologies is essential to understanding their capabilities and limitations (Camporeale et al., 2024).

This manuscript provided a timely opportunity to systematically explore the extent to which generative AI—specifically, OpenAI's ChatGPT—can assist in writing a scientific paper. The topic of this study is particularly well-suited for such an experiment: it does not introduce any novel methodology but rather applies well-established forecast verification techniques to publicly available data. The primary novelty of the work lies in the breadth of the validation effort—spanning 26 years of solar flare forecasts—which, to our knowledge, has not been previously conducted or benchmarked against standard baselines at this scale.

The experiment was conducted as follows: the first draft of the manuscript was authored by EC, who deliberately used ChatGPT (free version) as extensively as possible to generate the initial text. We estimate that approximately 80% of the initial draft was AI-generated or lightly edited by a human. The second author, TB, then revised and edited the text to ensure scientific accuracy, clarity, and coherence, without knowledge of which sections were generated by AI versus human-written. The resulting manuscript was then submitted for peer review in its final, collaboratively refined form. A similar exercise was then followed in addressing the reviewers' comments and preparing a revised version of the manuscript.

Our intent is to fully communicate to the journal readership which paragraphs have been AI-generated, so that everyone can judge by themselves the capability and limitations of generative AI in academic writing. In this experiment, the drafting and editing process was likely accelerated by a factor of two to three. Much of the accompanying software was also AI-generated and subsequently refined through human debugging and improvement. We found this especially valuable during the non particularly exciting phases of the project, such as data downloading, parsing, and cleaning.

The paper follows a *Neverending Story* style (Ende, 1993). The text in plain formatting has been AI-generated or only slightly edited (no more than a few words changed within a paragraph). The text in italics has been either entirely human-generated or severely edited starting from an AI text.

To promote transparency and reproducibility, we report in the Supporting Information S1 a list of prompts used during our chat with ChatGPT.

Data Availability Statement

The daily solar flare probability forecasts from NOAA/SWPC used in this study are publicly available at ftp://ftp.swpc.noaa.gov/pub/warehouse/. The solar flare event data prior to 2002 are available from ftp://ftp.swpc.noaa.gov/pub/indices/events/, while post-2002 data are accessed through the ASR flare catalogue from https://github.com/helio-unitov/ASR_cat/releases/download/v1.1/f_1995_2024.csv. All data used span the years 1996–2024 and are fully open access. A copy of all the data used in the paper, along with the code used for data pre-processing, analysis, and figure generation—including annotated Jupyter notebooks—is available on Zenodo at https://doi.org/10.5281/zenodo.16620803 (Camporeale, 2025) and in a public GitHub repository at https://github.com/ML-Space-Weather/solar_flare_verification. This ensures the full reproducibility of the results presented in this paper.

Acknowledgments
This work was partly inspired by a conversation between one of the authors and a representative from the Dutch Ministry of Defense, who shared that the NOAA/SWPC flare forecast was routinely used under the assumption that it was a regularly verified and skillful operational product. A subsequent literature search motivated the study published here. This work was partially supported by NASA under awards No 80NSSC23M0192, 80NSSC20K1580, 80NSSC21K1555.

References

Benz, A. O. (2017). Flare observations. *Living Reviews in Solar Physics*, *14*, 1–59. https://doi.org/10.1007/s41116-016-0004-3

Berretti, M., Mestici, S., Giovannelli, L., Del Moro, D., Stangalini, M., Giannattasio, F., & Berrilli, F. (2025). Asr: Archival solar flares catalog. *The Astrophysical Journal - Supplement Series*, *278*(1), 9. https://doi.org/10.3847/1538-4365/adc731

Camporeale, E. (2019). The challenge of machine learning in space weather: Nowcasting and forecasting. *Space Weather*, *17*(8), 1166–1207. https://doi.org/10.1029/2018sw002061

Camporeale, E. (2025). Code and data for swpc solar flare forecast analysis. *Zenodo*. https://doi.org/10.5281/zenodo.16620803

Camporeale, E., Marino, R., & Board, E. (2024). Our vision for jgr: Machine learning and computation. *Journal of Geophysical Research: Machine Learning and Computation* (Vol. 1). Wiley Online Library. https://doi.org/10.1029/2024jh000184. e2024JH000184.

Chen, Y., Manchester, W., Jin, M., & Pevtsov, A. (2024). Solar imaging data analytics: A selective overview of challenges and opportunities. *Statistics and Data Science in Imaging*, *1*(1), 2391688. https://doi.org/10.1080/29979676.2024.2391688

Crown, M. D. (2012). Validation of the noaa space weather prediction center's solar flare forecasting look-up table and forecaster-issued probabilities. *Space Weather*, *10*(6), S06006. https://doi.org/10.1029/2011sw000760

Deshmukh, V., Flyer, N., van der Sande, K., & Berger, T. (2022). Decreasing false-alarm rates in CNN-based solar flare prediction using SDO/HMI data. *The Astrophysical Journal - Supplement Series*, *260*(1), 9. https://doi.org/10.3847/1538-4365/ac5b0c

Ende, M. (1993). *The neverending story*. Penguin.

Florios, K., Kontogiannis, I., Park, S.-H., Guerra, J. A., Benvenuto, F., Bloomfield, D. S., & Georgoulis, M. K. (2018). Forecasting solar flares using magnetogram-based predictors and machine learning. *Solar Physics*, *293*(2), 28. https://doi.org/10.1007/s11207-018-1250-4

Georgoulis, M. K., Yardley, S. L., Guerra, J. A., Murray, S. A., Ahmadzadeh, A., Anastasiadis, A., et al. (2024). *Prediction of solar energetic events impacting space weather conditions*. Advances in Space Research.

Gopalswamy, N. (2018). Extreme solar eruptions and their space weather consequences. In *Extreme events in geospace* (pp. 37–63). Elsevier.

Hill, S., Pizzo, V., Balch, C., Biesecker, D., Bornmann, P., Hildner, E., et al. (2005). The noaa goes-12 solar x-ray imager (sxi) 1. instrument, operations, and data. *Solar Physics*, *226*(2), 255–281. https://doi.org/10.1007/s11207-005-7416-x

Hurlburt, N., Cheung, M., Schrijver, C., Chang, L., Freeland, S., Green, S., et al. (2012). *Heliophysics event knowledgebase for the solar dynamics observatory (sdo) and beyond* (pp. 67–78). The solar dynamics observatory.

Ji, A., Aydin, B., Georgoulis, M. K., & Angryk, R. (2020). All-clear flare prediction using interval-based time series classifiers. In *2020 ieee international conference on big data (big data)* (pp. 4218–4225).

Leka, K., & Barnes, G. (2018). Solar flare forecasting: Present methods and challenges. In *Extreme events in geospace* (pp. 65–98). Elsevier.

Leka, K., Park, S.-H., Kusano, K., Andries, J., Barnes, G., Bingham, S., et al. (2019). A comparison of flare forecasting methods. ii. benchmarks, metrics, and performance results for operational solar flare forecasting systems. *The Astrophysical Journal - Supplement Series*, *243*(2), 36. https://doi.org/10.3847/1538-4365/ab2e12

McIntosh, P. S. (1990). The classification of sunspot groups. *Solar Physics*, *125*(2), 251–267. https://doi.org/10.1007/bf00158405

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Pötzi, W., Veronig, A. M., Riegler, G., Amerstorfer, U., Pock, T., Temmer, M., et al. (2015). Real-time flare detection in ground-based Hα imaging at kanzelhöhe observatory. *Solar Physics*, *290*(3), 951–977. https://doi.org/10.1007/s11207-014-0640-5

Sadykov, V., Kosovichev, A. G., Kitiashvili, I., Oria, V., Nita, G., Illarionov, E., et al. (2021). "all-clear" prediction of solar proton events using machine learning and comparison with operational forecasts. In *Applications of statistical methods and machine learning in the space sciences*.

Schrijver, C. J., Kauristie, K., Aylward, A. D., Denardini, C. M., Gibson, S. E., Glover, A., et al. (2015). Understanding space weather to shield society: A global road map for 2015–2025 commissioned by cospar and ilws. *Advances in Space Research*, *55*(12), 2745–2807. https://doi.org/10.1016/j.asr.2015.03.023

Schwenn, R. (2006). Space weather: The solar perspective. *Living Reviews in Solar Physics*, *3*(1), 1–72. https://doi.org/10.12942/lrsp-2006-2

Temmer, M. (2021). Space weather: The solar perspective. *Living Reviews in Solar Physics*, *18*(1), 4. https://doi.org/10.1007/s41116-021-00030-3

Van der Sande, K., Flyer, N., Berger, T. E., & Gagnon, R. (2022). Solar flare catalog based on SDO/AIA EUV images: Composition and correlation with GOES/XRS X-ray flare magnitudes. *Frontiers in Astronomy and Space Sciences*, *9*, 1031211. https://doi.org/10.3389/fspas.2022.1031211

Zhao, L. (2023). Clear space weather center of excellence: All-clear solar energetic particle prediction. *arXiv preprint arXiv:2310.14677*.