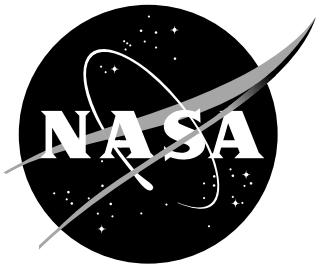
NASA/TP–20260000463



Validation of Solar Energetic Particle Forecasting Models for Space Radiation Operations with SPHINX and VIVID

Kathryn Whitman
KBR, Houston, TX

Ricky Egeland
NASA Johnson Space Center, Houston, TX

Clayton Allison
Leidos, Houston, TX

Philip Quinn
Leidos, Houston, TX

Luke Stegeman
Leidos, Houston, TX

January 2026

NASA STI Program . . . in Profile

Since its founding, NASA has been dedicated to the advancement of aeronautics and space science. The NASA scientific and technical information (STI) program plays a key part in helping NASA maintain this important role.

The NASA STI Program operates under the auspices of the Agency Chief Information Officer. It collects, organizes, provides for archiving, and disseminates NASA's STI. The NASA STI Program provides access to the NTRS Registered and its public interface, the NASA Technical Reports Server, thus providing one of the largest collections of aeronautical and space science STI in the world. Results are published in both non-NASA channels and by NASA in the NASA STI Report Series, which includes the following report types:

- TECHNICAL PUBLICATION. Reports of completed research or a major significant phase of research that present the results of NASA programs and include extensive data or theoretical analysis. Includes compilations of significant scientific and technical data and information deemed to be of continuing reference value. NASA counterpart of peer-reviewed formal professional papers, but having less stringent limitations on manuscript length and extent of graphic presentations.
- TECHNICAL MEMORANDUM. Scientific and technical findings that are preliminary or of specialized interest, e.g., quick release reports, working papers, and bibliographies that contain minimal annotation. Does not contain extensive analysis.
- CONTRACTOR REPORT. Scientific and technical findings by NASA-sponsored contractors and grantees.
- CONFERENCE PUBLICATION. Collected papers from scientific and technical conferences, symposia, seminars, or other meetings sponsored or co-sponsored by NASA.
- SPECIAL PUBLICATION. Scientific, technical, or historical information from NASA programs, projects, and missions, often concerned with subjects having substantial public interest.
- TECHNICAL TRANSLATION. English-language translations of foreign scientific and technical material pertinent to NASA's mission.

Specialized services also include organizing and publishing research results, distributing specialized research announcements and feeds, providing information desk and personal search support, and enabling data exchange services.

For more information about the NASA STI Program, see the following:

- Access the NASA STI program home page at ***http://www.sti.nasa.gov***
- E-mail your question to help@sti.nasa.gov
- Phone the NASA STI Help Desk at 757-864-9658
- Write to:
  NASA STI Information Desk
  Mail Stop 148
  NASA Langley Research Center
  Hampton, VA 23681–2199

This report is available in electronic form at
http://

Abstract

NASA human exploration missions seek to bring astronauts to new radiation environments outside the protection of the Earth's magnetosphere. In these environments, crew members experience an increased risk of radiation exposure from transient enhancements caused by solar energetic particle (SEP) events. SEP forecasting models seek to give operators both an assessment of the likelihood of these events and advance warning that can improve response times and be used in operational decision making. The ISEP project has been developed to gather and assess SEP forecasting models from across the field. In this report, we present the first systematic validation of model performance for a set of 11 models deployed on the ISEP SEP Scoreboards and validated in the SEPVAL community challenge. The validation results allow us to define state-of-the-art model performance and develop reasonable requirements for forecasting models to be used in operations. We also discuss the application of these results to understanding model performance from the operator perspective.

Contents

List of Figures

List of Tables

Executive Summary

The Space Radiation Analysis Group (SRAG) is tasked with the radiation protection of NASA astronauts, who are at an increased risk of radiation exposure from Solar Energetic Particle (SEP) events as Artemis missions take them into the deep space environment, outside the protection of the Earth's magnetic field. Reliable forecasts of these events would enable SRAG console operators to expedite protective actions for deep-space crews. The Integrated Solar Energetic Proton Event Alert/Warning System (ISEP) project, a collaborative effort between NASA's SRAG, Community Coordinated Modeling Center (CCMC), and Moon to Mars Space Weather Analysis Office (M2M), has aggregated and improved SEP forecasting models from throughout the community, and has produced the SEP Scoreboards as a single system for forecasts to be visualized and monitored by SRAG console operators in real-time. This system has allowed—for the first time—a comprehensive quantitative validation of forecasts issued by SEP models in real time. In parallel, a dedicated community validation challenge, SEP Model Validation Working Meeting (SEPVAL), solicited model forecasts for a benchmark set of SEP event and non-event periods, allowing for model evaluation and comparison on a consistent basis.

The SEP model forecast landscape is extremely complex, with dozens of models forecasting a subset of several characteristics of SEP events, using different observables as inputs, forecasting for different particle energy channels, and operating with widely different cadence and prediction windows. To evaluate models in this complex landscape, SRAG has developed the Solar Particles in the Heliosphere validation INfrastructure for SpWX (SPHINX) Validation Framework to prepare satellite observations as ground truth, automatically match SEP model forecasts to observations, calculate a wide variety of validation performance metrics, and finally generate reports for end users. Additionally, Validation in Visually Interactive Displays (VIVID) was developed as a stand-alone interactive web tool that allows users to investigate SPHINX results by applying filters, generating plots and recalculating metrics on the fly. SPHINX was developed to be robust to the unique challenges of real-time forecasting, including duplicate, incomplete, or corrupted forecasts, the ability to incrementally update validation results with new forecasts, and optimization to manage the hundreds of thousands of forecasts issued to the Scoreboards each month.

This technical report evaluates a set of 11 models that have been operating in real time on the Scoreboard since as early as March of 2020, and have participated in the SEPVAL challenge. The goal of the evaluation is to develop a quantitative understanding of model performance to communicate to console operators, to compare models to one another, to inform and update SRAG's model requirements, and to assess the current state of the art in SEP forecasting and where this capability stands with respect to our needs as stated in the previous NASA gap analysis. It is important to note that this report presents validation results that represent SEP model forecasting performance with respect to SRAG's operational needs. Model forecasts are compared to >10 MeV and >100 MeV operational GOES proton measurements provided by NOAA as these are the particle fluxes used by SRAG to

monitor the radiation environment for astronaut protection. This report is the result of a large-scale numerical analysis of over 8 million model forecasts covering 10 validated quantities evaluated according to more than a dozen metrics. We have found that model performance is multi-faceted and cannot be adequately expressed by a single metric. Nonetheless, every effort has been made to simplify the validation to the largest degree possible.

SPHINX produced metrics for many forecasted quantities, but here we focus on All Clear, Probability of Occurrence, and peak flux predictions. All Clear is defined as a binary forecast of "Clear" or "Not Clear" indicating whether the particle intensity is expected to remain below or exceed a threshold, e.g., >10 MeV protons exceeding 10 pfu.

General conclusions regarding the validation and model performance are made, followed by general recommendations intended to contribute to the Operations-to-Research process in the R2O2R cycle.

Validation and model performance outcomes:

1. Models have some ability to discriminate between solar conditions that produce >10 MeV SEP events and those that do not. None of the models evaluated here had skill in forecasting >100 MeV events.

2. For real-time All Clear forecasts on the SEP Scoreboards, event Hit Rates show some skill (∼50–70% median values) for >10 MeV forecasts, but are very poor for >100 MeV (18% median for flare/CME post-eruptive models).

3. For the SEP Scoreboards, a "Not Clear" forecast turns out to be clear in the vast majority of the cases, >70% median False Alarm Ratio across models. The false alarm problem is substantially worse for >100 MeV forecasts than >10 MeV forecasts, and worse for regular cadence pre-eruptive forecasts than for post-eruptive models triggered by flares or Coronal Mass Ejections (CMEs).

4. For the imbalanced Scoreboard dataset, we find a median HSS of 0.28 for >10 MeV forecasts across all flare/CME post-eruptive models, and 0.03 for pre-eruptive forecasting models. Median values increase to 0.47 for post-eruptive models and 0.07 for pre-eruptive models on the balanced SEPVAL dataset.

5. The highest skill models for >10 MeV All Clear on the Scoreboard were UMASEP-10 with a Hit Rate of 69%, False Alarm Rate of 3%, False Alarm Ratio of 63%, and HSS of 0.46. SEPSTER achieved the second highest HSS of 0.44 with a Hit Rate of 62%, False Alarm Rate of 1.8%, and False Alarm Ratio of 64%. Comparing HESPERIA REleASE SOHO-60 min's warning conditions for 15.8–39.8 MeV with GOES >10 MeV threshold crossings results in a Hit Rate of 88%, False Alarm Rate of 0.1%, False Alarm Ratio of 42%, and HSS of 0.70.

6. NOAA SWPC Warnings have high skill with a 100% Hit Rate and 29% False Alarm Ratio for the SEP Scoreboards and an 85% Hit Rate and 3% False Alarm Ratio for SEPVAL. Thus, the Warnings surpass the skill of all models

while achieving a median advance warning time of an hour, providing a clear benefit to operations.

7. For the SEP Scoreboard period, NOAA SWPC has a more skillful 24-hour >10 MeV probability forecast than any of the Scoreboard models when analyzed over 5 years of real solar conditions ranging from quiet to active, but in the SEPVAL challenge dataset its performance was worse than the models. Considering the bias toward complex active regions that produced strong flares and fast CMEs in the SEPVAL dataset, this could be an indication that Space Weather Prediction Center (SWPC)'s ability to discriminate such conditions between events and non-events is worse than models performing numerical analyses of active regions.

8. Pre-eruptive forecasting models (e.g. MAG4, MagPy) predicting the probability of occurrence for >10 MeV events show a limited ability to distinguish event periods from non-event periods. In the best case (MAG4 LOS_FEr), a 3% increase in the median probability issued preceding SEP events is seen.

9. Post-eruptive forecasting models (e.g. SEPSTER2D, SEPMOD) show limited success in predicting the peak flux. Most predictions are within a factor of 10 and $\sim$1/3 are within a factor of 2 of the observed value, but errors extend out to multiple orders of magnitude. Peak flux forecasting performance significantly deteriorates for >100 MeV events. Overall, models can generally reproduce the statistical trending between large and small SEP events, but cannot predict the intensity of any particular event with accuracy.

10. CME-triggered post-eruptive models generally do not provide advance warning prior to observed threshold crossings for >10 MeV or >100 MeV events, but can provide peak flux estimates in advance.

11. Models taking *in situ* energetic proton or electron fluxes as input (UMASEP and REleASE) have significantly lower occurrences of false alarms and provide useful advance warning. UMASEP gives median advance warning on the order of 45 minutes for >10 MeV events and 20 minutes for >100 MeV events. HESPERIA REleASE provides 1 to 3.5 hours median advance warning for >10 MeV events, depending on the observational data stream. These models' ability to issue a forecast minutes to hours ahead of an SEP event, coupled with a relatively high level of skill, make them potentially useful models for operations.

General recommendations for model improvement informed by validation results:

1. Previous modeling efforts have focused on the NOAA SWPC definition of an SEP event in the >10 MeV energy channel, however space radiation operations are primarily concerned with energetic events that show enhancements for >100 MeV particles. Significant improvements are needed to forecast energetic >100 MeV SEP events and concentrated efforts towards that goal is identified as a priority for future model development.

2. The combination of mediocre Hit Rates with high False Alarm Ratios make it difficult to trust model All Clear forecasts in operations. Development of strategies to achieve high Hit Rates while reducing false alarms is identified as a priority to increase reliability for operations.

3. For All Clear forecasts, the number of false alarms tends to be much higher than the number of observed SEP events in real-time operations. To achieve high skill in an imbalanced, climatological scenario, False Alarm Rates must be on the order of a few percent while False Alarm Ratios should be less than 50% such that the false alarms do not outnumber hits.

4. The Heidke Skill Score (HSS), widely used in SEP forecasting model validation, is extremely sensitive to dataset imbalance and is low when the False Alarm Ratio is high on imbalanced datasets. The use of numerical values for this metric as a requirement must be accompanied by conditions on the dataset for which it is measured.

5. HSS is sensitive to both the Hit Rate and False Alarm Ratio, making it a good choice for optimization of model performance (e.g., for development of machine learning models) compared to other widely used metrics such as True Skill Score (TSS). Since the numerical value of HSS is sensitive to imbalance in the dataset, a numerical target has limited utility, but optimization of HSS for a given dataset should result in skill that is calibrated towards space radiation operations needs.

6. Validation results for probability forecasts show that there are simple modifications that could improve aspects of model performance. In particular, inspection of ROC curves (see Section 3.3.2 for ROC definition) for models forecasting event probability reveals that there is room for tuning and optimization of Hit Rate/False Alarm Rate by choosing an alternative probability threshold.

7. Pre-eruptive forecasting models (e.g., MAG4, MagPy) have a very challenging task to discern between active region conditions that will produce only flares and conditions that will produce both flares and SEPs. However, these are the only types of models that are likely able to provide significant advance warning, if they gain skill. Studies to identify new or supplemental pre-eruptive parameters with predictive power is needed.

8. CME-based models have demonstrated skill, but due to delays in receiving the necessary CME parameters, they struggle to produce forecasts on a timeframe useful for operations. Any reduction of time to input CME parameters into the models would improve advance warning. Tools, data availablity, or other ways to speed up CME measurements would be beneficial, e.g. high cadence coronagraphs, low latency data streams, or automated extraction of CME parameters.

9. Accuracy of peak flux forecasts must be increased while scatter is decreased in order for peak flux forecasts to become reliable enough for operational use. This implies that the conditions responsible for event-to-event variability must be better understood and reproduced by models. The paradigm must shift from predicting statistically average outcomes to predicting the outcome for **this** event.

10. Models should optimize on continuous metrics, as well as categorical (i.e. HSS). For example, reducing mean log error for peak flux models. Using continuous metrics reduces the error between observations and predictions, improving accuracy without assuming specific SEP event definitions or thresholds.

11. **In light of operational needs, SRAG proposes a new skill score called False Alarm Event Ratio (FAER)**, pronounced "fear", defined as: $FAER = false\ alarms/(hits + misses)$. This ratio represents the number of false alarms compared to the number of observed SEP events and has a simple intuitive utility for communicating performance to operators as models with high FAER are impossible to trust.

1 Introduction

The Space Radiation Analysis Group (SRAG) at NASA Johnson Space Center (NASA JSC) continuously monitors the space weather environment with the aim to protect astronauts from enhancements in space radiation. Eruptions in the solar corona, typically flares and coronal mass ejections (CME) are associated with the release of solar energetic particles (SEP), charged particles mainly consisting of protons that travel at near-relativistic speeds. These charged particles can arrive at Earth as quickly as 20–30 minutes following an eruption where they can then penetrate satellite hardware or the shielding of crewed vehicles and cause radiation impacts to electronics and humans in space (Hu et al., 2009).

For the last two decades, SRAG has supported space radiation operations for astronauts onboard the International Space Station (ISS). The ISS orbits at an altitude of 250 km, well within the Earth's protective magnetic field, which acts to deflect most SEPs that arrive at Earth. The ISS is typically vulnerable to SEP enhancements only during short (10–20 minute) orbital passes near the geomagnetic poles, naturally reducing the radiation risk.

The Artemis missions, starting with Artemis II, will take astronauts beyond the Earth's magnetic field, exposing them to the full duration of an SEP event should one occur. In this scenario, SRAG must respond quickly to solar activity and potential changes in the space radiation environment to determine whether astronauts should take action to reduce radiation exposure or conclude that no action is required. In support of this goal, the ISEP project was established to investigate the use of SEP forecasting models in SRAG console operations. ISEP is a collaboration between NASA's SRAG, NASA's CCMC, and NASA's M2M. Through this collaboration, the SEP Scoreboards[1] have been developed to visualize forecasts from SEP prediction models running in real time, as well as official alerts and warnings from the NOAA SWPC. SEP model forecasts are currently being monitored by SRAG operators to evaluate their utility for Artemis operations.

Over the last two decades, measurements of the space weather system have proliferated, ranging from remote sensing observations of the Sun to *in situ* measurements at Earth. Near real-time observations of photospheric magnetic fields, extreme ultraviolet (EUV) images of coronal structures, X-ray measurements of flares, white-light coronagraph images of solar wind structures and coronal magnetic fields, radio signatures of particle acceleration, *in situ* magnetic field measurements of the solar wind and interplanetary shocks, and *in situ* energetic particle measurements provide a wide variety of resources for understanding and predicting the space weather environment. In addition, new Machine Learning (ML) and Artificial Intelligence (AI) techniques have become popular in the research community, while computational power has increased significantly, making computationally-intensive physics-based models more accessible. This has led to the development of a wide variety of model approaches that take advantage of different combinations of available model inputs and produce many different forecast outputs.

With the proliferation of SEP forecast models, the development of real-time

[1] https://ccmc.gsfc.nasa.gov/scoreboards/sep/

forecast tools through ISEP, and the need for a fast response to changing space weather conditions during Artemis, it has become imperative to quantify individual model performance and define the state of the art of SEP model forecasting. A generalized validation infrastructure is needed to evaluate the wide array of SEP model approaches and predictions. SPHINX was created to meet this need.

The SPHINX Validation Framework was developed through a two-pronged approach — a validation challenge within the research community for a benchmark set of events and the validation of the continuously-running, real-time SEP Scoreboards. Both approaches have been highly successful Reasearch-to-Operations-to-Research (R2O2R) efforts that have led to new understanding of SEP model performance while inspiring improvements and new developments in many of the participating models. At the same time, these efforts have enabled the development of SPHINX into a generalized, robust, and flexible tool.

SRAG and ISEP have been leading an SEP model validation effort within the research community for more than six years that defined a set of challenge events and solicited forecasts from all interested model developers, regardless of model type, maturity, or underlying approach. This challenge was carried out through the Solar Heliospheric and INterplanetary Environment workshop (SHINE), Committee on Space Research (COSPAR), and ESWW conferences. An International Space Weather Action Team (ISWAT) was created for SEP model validation (H3-01)[2]. Additionally, two dedicated SEPVAL working meetings were organized in 2023, one in the US and one in Europe, that focused on the progress of the validation effort and facilitating R2O2R communication. Topics included:

- Progress reports of the SPHINX validation infrastructure with the goal to reach community concurrence and consensus
- Reporting validation results to and soliciting feedback from model developers
- Communication between researchers and representatives from NASA SRAG and CCMC to relay how models could be onboarded into the SEP Scoreboards at CCMC and contribute to SRAG operations
- Communication between researchers and representatives from NOAA SWPC and the ESA Space Weather Office to clarify the R2O process into space weather operations

Starting in 2019, the first community challenge defined a small set of ten "SHINE challenge" SEP events and invited the community to submit forecasts. This challenge list was later expanded to 33 SEP events and a nearly equal number of 30 non-event (quiet) periods to enable a more complete validation of each model. The community challenge played a critical role in the development of SPHINX as a generalized tool. It allowed SRAG to survey the research community and better understand the models and predictions under development. SPHINX had to be developed to be robust to the range and combinations of forecasts. Workflows had to be developed that validated the *intent* of each model. To ensure that metrics

[2] https://iswat-cospar.org/H3-01

typically reported by the research community were included, a wide variety of metrics were programmed into SPHINX. In reviewing the validation results, it became evident that no single metric accurately represented model performance, but rather a set of metrics was required to build a full picture, inspiring the further expansion of SPHINX's metrics library. Finally, the community challenge enabled progress in the development of a validation framework while the SEP Scoreboards were still under development.

ISEP goals are to identify, transition, and evaluate new models (R2O); develop CCMC SEP Scoreboard software tailored for SRAG; and implement these capabilities within CCMC as a non-operational prototype. The ISEP group worked with model developers to implement the capability to run models robustly in real time and, in some cases, to improve or expand models to provide additional predictions useful to SRAG. The SEP Scoreboards were developed by NASA CCMC with significant input from SRAG to ensure that they provide relevant information in a useful format for SRAG console operators. Shortly after the beginning of the ISEP project, M2M was established. One element of the ISEP partnership is the transitioning of ISEP models/software from CCMC to M2M. M2M hosts the SEP models and SEP Scoreboards within its own production environment, serving as both a proving ground and operational framework. M2M supports the evaluation of SEP models with real-time analysis of the space weather environment provided to SRAG console operators. M2M provides critical human-in-the-loop activities required to run the SEP Models and SEP Scoreboards in real time.

Model validation results published in the literature are calculated using historical event lists that may not accurately represent the real climatology of SEP events and non-event periods. In these types of studies, models may be trained on data that occurs in the future of a given "test" event. Furthermore, historical analyses do not typically include the real-time availability of model inputs. For these reasons, the SEP Scoreboards are extremely valuable tools for collecting genuine forecasts with no knowledge of the future using the true availability of real-time data sources, demonstrating model performance in an operational-like scenario.

Validation of the SEP Scoreboards comes with its own set of challenges for SPHINX. Many of the models on the SEP Scoreboards produce forecasts every few minutes or with an hourly cadence. There may be dozens of flares or CMEs within a short time period, triggering forecasts for each occurrence. Together, these result in millions of forecasts over the lifetime of the Scoreboards, introducing computational challenges. Further complications include forecast revisions and duplication. SPHINX had to be developed to be robust to each of these issues.

Through the two developmental pathways, SPHINX has reached a high level of maturity. In this report, we use the SPHINX Validation Framework to derive first results for SEPVAL and the SEP Scoreboards and define the state of the art in SEP model forecasting. A description of the framework is outlined in Section 2. Details about the participating models and validation data sets are described in Section 4. Summary results reporting group median metrics and range of performance for each effort are presented in Section 5. Detailed descriptions of individual model performance are provided in Section 6. Many models in the research community

participated in the SEPVAL effort; however, in this report we focus on the models that are active in the SEP Scoreboards as they are currently being used by SRAG for situational awareness during console operations. In Section 7, cross-model comparisons of similar models investigate the performance of different implementations of similar concepts. Section 8 highlights major take-aways and provides a definition of the state-of-the-art SEP model performance. In Sections 8.3 and 8.3, we relate the model results described here to NASA gaps and SRAG's requirements. We provide final thoughts in Section 8.4.

Figure 1.1: SPHINX logo.[†]

Figure 1.2: SEPVAL logo.[†]

Figure 1.3: VIVID logo.

[†]Artwork by Jordan Stegeman, used with permission.

2 Validation Infrastructure

SPHINX: A gatekeeper who devours all who do not correctly answer her riddle.

The SPHINX Validation Framework consists of a set of programs that prepare observational inputs, organize forecasts, evaluate model performance, and display the performance results. SPHINX, an automated, generalized validation program developed to evaluate forecasts from SEP prediction models, forms the heart of the Framework. The term SPHINX will be used interchangeably throughout this report to represent the validation program and the broader framework.

The SPHINX Validation Framework, outlined in Figure 2.1, is comprised of the individual programs:

- fetchsep
  - Prepare observed "truth" values from satellite particle measurements
  - Prepare forecast jsons for models that output time profiles
  - Manage lists of observation files
- fetchcasts
  - Download forecasts submitted to the SEP Scoreboards from CCMC's iSWA servers
  - Manage lists of forecast files
- sphinxval (SPHINX)
  - Automatically associate (match) observed values with forecasted values
  - Calculate metrics and produce plots
  - Generate HTML reports for individual models and summary plots for model cross-comparisons
- mailsphinx
  - Generate a periodic report of SEP forecasting model activity, including validation information, and emails information to subscribers (model developers, console operators, management, etc.)
  - Summarize forecast statistics over the summary period, year-to-date, and all-time
  - Summarize space weather during the summary period, with special attention paid to any solar particle events that occurred

- VIVID
  - Web application for displaying validation results in a dashboard of interactive plots and tables
  - Apply filters (e.g., time range, CME speed, etc.) and recalculate metrics

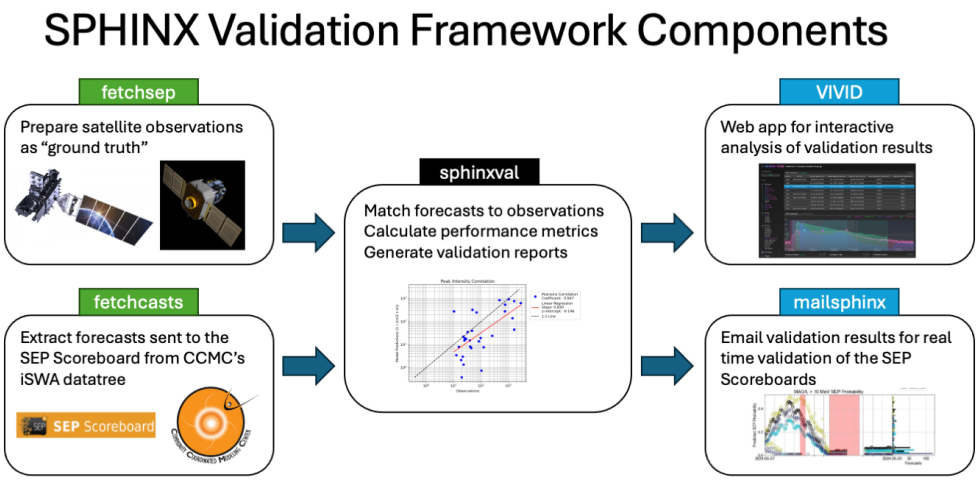


Figure 2.1: The main components that make up the SPHINX Validation Framework. Taken together, the components create an end-to-end workflow to prepare observations, receive forecasts, automatically associate forecasted and observed values, calculate metrics, and report metrics in a variety of formats and interactive tools.

2.1 FetchSEP

The FetchSEP package contains two distinct programs:

- IDSEP - processes long timescale particle intensity time series and automatically produces SEP event lists
- OpSEP - processes individual SEP events, extracting the SEP event characteristics and saving in the appropriate JavaScript Object Notation (JSON) format

FetchSEP first runs IDSEP to identify all enhancements above background due to SEP events over a long timescale, e.g., a solar cycle or more. Figure 2.2 shows an example of enhancements above background identified by IDSEP as SEP events (orange). The SEP list is then processed by OpSEP, which creates individual JSON

files for each SEP event and the quiet time periods between them. OpSEP calculates SEP event start, end, and peak times, peak values, and event-integrated fluence spectra. In this way, the JSON files produced by FetchSEP contain the ground truth values for every moment in time for a continuous observational period. Figure 2.3 shows the values calculated by OpSEP for an SEP event in November 2011.

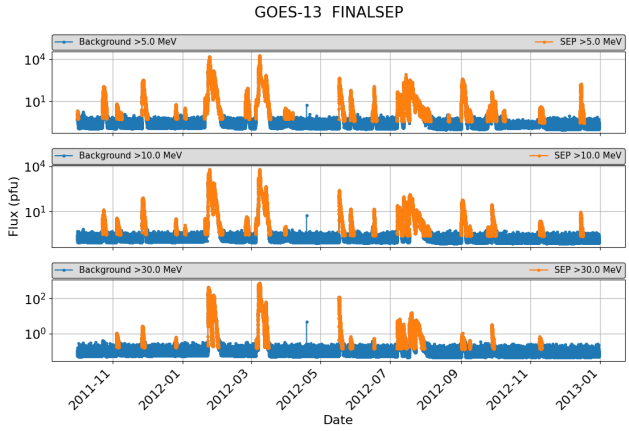


Figure 2.2: Automatic identification of proton enhancements due to SEP events (orange) using IDSEP.

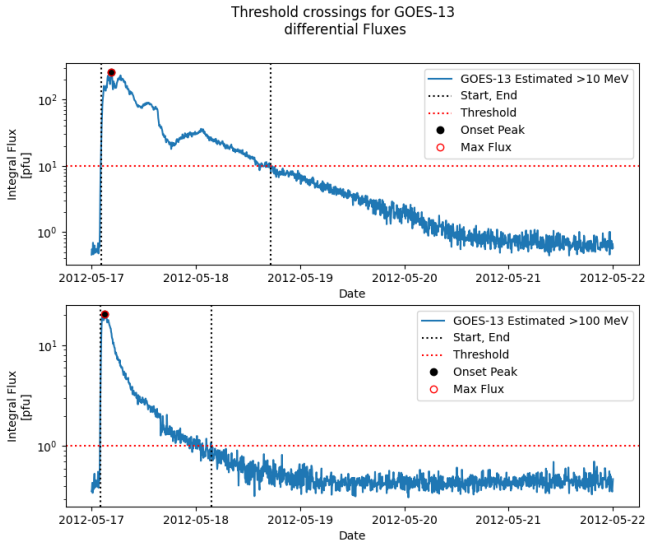


Figure 2.3: Derivation of various quantities for a specific SEP event on May 17, 2012 using OpSEP. A threshold of 10 pfu is applied to >10 MeV and 1 pfu is applied to >100 MeV to created event definitions.

OpSEP calculates two types of peak flux values, which SPHINX validates separately. The "onset peak" is the point at which the initial particle rise stops or slows,

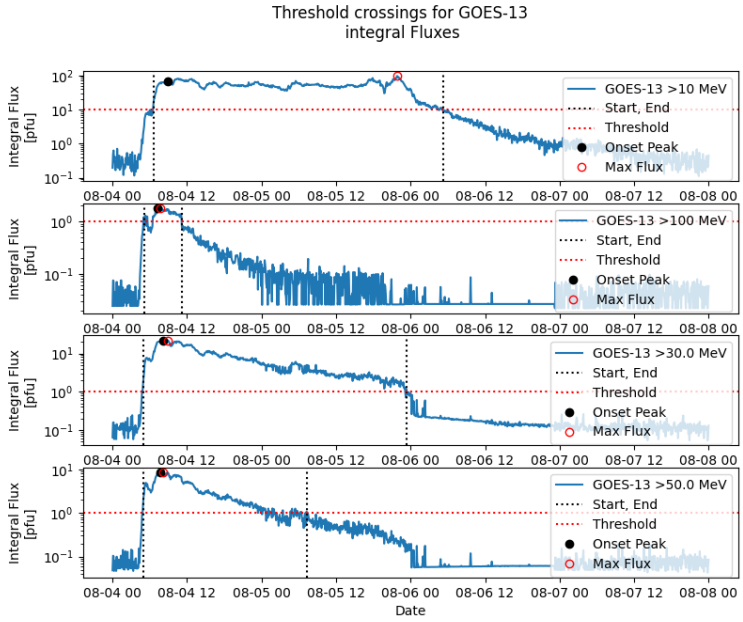


Figure 2.4: The black circle indicates the "onset peak", the point at which the particle intensity finishes its initial rise. The red circle indicates the "max flux", the maximum flux achieved during the SEP event.

shown as the black circles in Figures 2.3 and 2.4. The onset peak is most closely related to the particle acceleration physics in the low solar corona and is typically the peak predicted by first-principles physics-based models. The "max flux" is the maximum flux achieved during the full SEP event. This peak may be the same as the onset peak, as seen in Figure 2.3, or very different, as seen in Figure 2.4. The maximum flux depends on particle transport effects and CME passages that may cause Energetic Storm Particle (ESP) enhancements, particularly for energy channels below 30 MeV.

The max flux is a straightforward value to calculate and reproduce. The onset peak is more difficult and subjective to define. FetchSEP takes the approach that all derived SEP quantities should be fully reproducible and therefore applies an automated algorithm to identify the onset peak. Using the threshold crossing time as a reference point, the algorithm extracts a portion of the time profile a number of hours prior to the threshold crossing and up to 18 hours after in order to capture the initial SEP rise. A Weibull function is fit to this portion of the profile, shown in Figure 2.5 (orange). The maximum of the first order derivative of the fit taken with time identifies the location of the maximum of the Weibull fit (green circle) and a minimum in the second order derivative identifies the initial turning point of the fit (purple diamond). The second derivative was found to identify the onset peak more effectively. The final value of the onset peak (brown triangle) is selected as the maximum measured flux within ± 1 hour of the minimum in the second derivative.

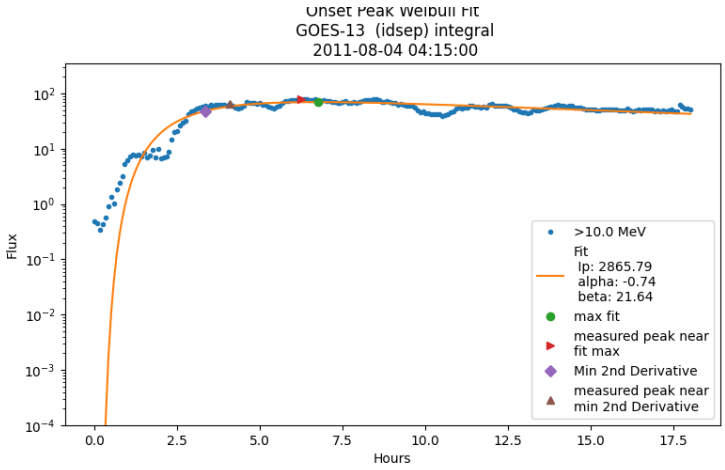


Figure 2.5: To identify the onset peak, a Weibull function is fit to the rising portion of the SEP event (orange). A maximum in the first order derivative of the fit (green circle) and a minimum in the second order derivative of the fit (purple diamond) were investigated as approaches to identify the onset peak. The final onset peak (brown triangle) is taken as the maximum measured flux within ±1 hour of the minimum in the second derivative.

OpSEP processes individual SEP event time profiles to extract SEP event information, including time profiles produced by SEP models. OpSEP may be used to read in a predicted time profile, extract all predicted parameters, like threshold crossings, peak flux, and fluence spectra, then export the prediction in the CCMC JSON format for the SEP Scoreboards and SPHINX.

2.2 FetchCasts

The fetchcasts package manages the download and synchronization of forecast JSONs from the CCMC SEP Scoreboard (or integrated Space Weather Analysis System (iSWA) Data Tree) to a local directory. FetchCasts works on a per-model and per-month basis to download the forecast data from the iSWA Data Tree. It uses custom per-model file filters to avoid downloading data that is not needed by SPHINX. The program evaluates the local forecast directory and can generate forecast JSON lists to be used by SPHINX on a per-model and per-month (or time period) basis. Additionally, the program can produce statistics on the number of forecasts per model and over time, that can be used to understand the forecast cadence and the overall workload for SPHINX.

2.3 SPHINX

SPHINX (i.e., the sphinxval package) performs three major functions:

- Automatically identifies the appropriate observed “truth” values to compare to forecasted values, a process referred to as “matching”

- Calculates metrics from the paired observed and forecasted values
- Generates output files, plots, and reports

SPHINX is meant to be used for both historical validation of discrete time periods and for the validation of the SEP Scoreboards which continuously receive SEP forecasts in real time. Some models produce forecasts every few minutes, resulting in millions of forecasts over the lifespan of the Scoreboards. This makes it infeasible to do a manual comparison between observations and forecasts, driving automation of every step in the validation process.

In general, SPHINX identifies the appropriate observational values to be compared to each forecasted value. Referring to the CCMC SEP Scoreboard JSON schema [3], any of the following forecasted values may be present in a forecast JSONs, thus the observed value of each is determined:

- All Clear
- Probability
- Threshold Crossing Time
- Start Time
- End Time
- Peak Flux
- Peak Flux Time
- Fluence
- Fluence Spectrum
- Time Profile

When performing automated "matching" of observed to forecasted values, the algorithm uses the information prepared from the observed proton intensity time series by FetchSEP along with timing of the forecast inputs/triggers and the prediction window specified in the forecast JSONs.

In general, SPHINX takes the approach that a model prediction is considered a forecast if *all data input into the model is earlier in time than the observed phenomenon.* If a forecast is triggered by an eruption (flare or CME), additional checks of the timing are done to determine whether that specific eruption should be associated with a particular observed SEP event. Note that SPHINX does not require a forecast to be issued (made available to forecasters) before the observed phenomenon, discussed further below.

The matching and validation procedure proceeds via the following workflow:

- Each forecast undergoes a check that requires that the issue time is the same or later than the last data time of the inputs/triggers. A valid forecast cannot be issued prior to the last data ingested to produce the forecast.
- The forecasts are sorted by model, energy channel, and threshold.
- For each forecast, SPHINX identifies the observation files that contain measurements in the time period specified by the forecast prediction window.
- If no observed SEP event occurs in the forecast prediction window, then observed values are set accordingly.

[3] https://ccmc.gsfc.nasa.gov/scoreboards/sep/#submission-file-format-information

- If an observed SEP event occurs in the forecast prediction window, then a number of checks are performed related to the timing of the forecast.
- Once the matching criteria have been determined, a revision step is performed. For models that use flare/CME inputs, if multiple forecasts triggered by different flares/CMEs are matched to the same observed SEP event, then SPHINX applies simple logic to identify the most likely eruption as the cause of the SEP event and "unmatches" the other less likely forecasts (see Section 3.2.7 for more detail).
- The final outcome of the matching logic is stored as a Match Status, which is used later by the validation module.
- Finally, the observed and forecasted values and matching information are passed on to the validation module where metrics and plots are calculated and saved in output files.

SPHINX calculates a wide variety of metrics that represent accuracy, precision, correlation, and skill. As highlighted in (Liemohn et al., 2021), multiple metrics are needed to describe a complete picture of model performance. SPHINX calculates nearly all metrics used in the model community, allowing users to select their preferred subset of results. See Appendix A for the complete set of metrics calculated for each quantity.

In the case that models generate multiple forecasts for a single SEP event, SPHINX provides metrics for subsets of forecasts, i.e., first, last, mean, and max, to enable interpretation of model performance from different perspectives. For example, the metrics for the "first" forecast subset are derived from the first forecast that was made for each SEP event, giving insight into the performance of the earliest forecasts available to operators. Likewise, the "last" forecast subset describes model performance with the most amount of information available to the model prior to the SEP event. For probability or peak flux models, it is interesting to look at the maximum value provided for each SEP event.

SPHINX evaluates model performance and forecast availability separately; it takes the approach that model output is considered a forecast if all data input into the model is earlier in time than the observed phenomenon. Performance is calculated for all forecasts that satisfy this criteria. The speed at which the model is able to produce the forecast or the availability of the forecast in real time is determined by the issue time and is calculated as a separate metric, the Advance Warning Time (AWT), defined as the time between the issuance of a"Not Clear" forecast and the observed SEP event start time. Model performance and AWT taken together indicate utility in an operational setting. With this approach, SPHINX may be used both for scientific and operational validation purposes.

A number of features are implemented in SPHINX to manage the challenges of forecasting in real time to the SEP Scoreboards, namely large data sets, corrupt and duplicate forecasts, and adjustments to human analyses.

Handling the millions of forecasts generated over the lifetime of the Scoreboards requires optimization of SPHINX to speed up processing and reduce unnecessary

memory usage to ensure that the program can run in a reasonable amount of time without exceeding the available computational resources. The "resume" function processes forecasts in weekly or monthly increments, cumulatively adding them to previously calculated results until the complete set of forecasts is evaluated. This procedure allows SPHINX to add new forecasts submitted to the SEP Scoreboards to the validation metrics without having to reprocess the entire history of the SEP Scoreboards each time.

SPHINX applies duplicate forecast checking at many stages throughout the validation workflow to ensure that only unique forecasts are included in the final validation results. Duplicate forecasts are produced in the SEP Scoreboards for a variety of reasons. During data gaps, many models continue to produce forecasts using the last available data as input even though the data has become stale, creating many duplicate forecasts with no new information. These are not considered new forecasts in the SPHINX philosophy and are not be counted in the validation metrics. Some models implemented on the SEP Scoreboards have produced duplicate forecasts due to a bug. For example, during one particular month SEPSTER experienced a bug that caused the model to issue duplicate forecasts every 5 minutes until the hard drive of the CCMC host computer was out of space. The bug was identified and corrected. The duplicate forecasts have since been deleted, but at the time of processing the results presented in this report, they were still present and had to be identified and removed by SPHINX. To handle these cases and more, duplicate checking was implemented in SPHINX at many points within the workflow to ensure that only unique forecasts are included in the final validation results.

M2M uses coronagraph imagery from Geosynchronous Orbit Earth observing Satellite (GOES), Solar and Heliospheric Observatory (SOHO), and Solar TERrestrial RElations Observatory (STEREO) as it becomes available to make CME measurements which are then entered into CCMC's Database Of Notifications, Knowledge, Information (DONKI). M2M will typically first make measurements with as few images as possible so that models that need these CME values can produce a forecast as quickly as possible (see Table 4.8). As more images become available, M2M will make new measurements and enter these higher-quality CME parameters into DONKI which automatically trigger revised forecasts. SPHINX applies logic in the observation-to-forecast matching step that makes use of flare and CME information. This logic is developed to ensure that forecasts from the same flare or CME, even with slightly different parameters, are evaluated in the same way.

More details about SPHINX functionality is described in Section 3.2

2.4 MailSPHINX

MailSPHINX is a program that generates periodic reports of SEP forecast model validation metrics and sends these reports to interested parties (model developers, radiation console operators, etc.) via email. A typical MailSPHINX report is organized into a few sections:

- an overview section that notes the total number of SEP forecasts produced during the summary period, year-to-date, and all-time,

- an event section containing a list of SEP events that occurred during the summary period (if any), alongside event timing statistics,
- an all-clear contingency table for each forecasting model under consideration, including the number of hits, misses, false alarms, correct negatives, total forecasts (all-time and submitted during the summary period), and links to SPHINX validation reports,
- a summary of relevant space weather during the summary period (GOES X-ray flux, ACE differential electron flux, GOES integral proton flux), and
- a model performance section, which includes a series of plots that show model outputs as a function of time.

MailSPHINX is intended to run at a set cadence (e.g., every Monday at 00:00 UTC) immediately after a periodic SPHINX execution. Validation metrics are collected from SPHINX outputs, space weather timelines are collected from SWPC, and those data are ultimately compiled into a series of email-compatible graphs, tables, and text prior to distribution. Figure 2.6 shows an example of a potential MailSPHINX email format.

2.5 VIVID

VIVID is a web app developed at SRAG and hosted at the CCMC (https://web-dev.ccmc.smce.nasa.gov:8001/vivid/, note: a NASA VPN is currently required). There are plans to make VIVID available publicly in the near future. VIVID displays SPHINX results in a series of tables and plots, allows users to apply filters, and immediately recalculates all metrics. An example view of VIVID is shown in Figure 2.7. VIVID serves the purposes listed below and detailed in the paragraphs to follow.

- The SRAG validation team can further explore SPHINX results without the need to rerun SPHINX
- SRAG console operators can filter SPHINX results to check model performance for current radiation environment conditions
- The SRAG validation team can use VIVID to compare models side-by-side and determine the state-of-the-art performance in order to drive future work plans
- Model developers can independently explore SPHINX results and determine if their model has any internally-correlating conditions
- SRAG, model developers, the space weather community can download plots and tables for publications and presentations, or even use VIVID in real-time at a conference or working meeting

MailSPHINX Report

Report Generation Time: 2024-10-02 16:17 (all UTC)

Evaluation Period: 2024-05-01 00:00 to 2024-05-15 00:00

← MailSPHINX Archive

Overview

Time Period	Forecasts	Not Clear Forecasts	Above Threshold Peak Flux Forecasts
This Period: Since 2024-05-01 00:00	17364	6273	1630
This Year: Since 2024-01-01 00:00	262520	27083	5858
All Time: Since 2017-09-04 23:05	3014028	82481	18476

Event Summary

Energy [MeV]	Flux Threshold [pfu]	Observatory	Threshold Crossing Time	End Time	Duration [hr]	Fluence [cm^{-2}]	Onset Peak Flux [pfu]	Onset Peak Time	Max Flux [pfu]	Max Flux Time
≥ 10	≥ 10	GOES-18	2024-05-10 14:05	2024-05-11 01:00	10.9167	N/A	35.7863	2024-05-10 15:10	206.9193	2024-05-10 17:45
≥ 30	≥ 1	GOES-18,GOES-18	2024-05-10 17:45	2024-05-11 01:00	7.2500	N/A	0.8740	2024-05-10 16:45	1.2423	2024-05-10 17:55
≥ 10	≥ 10	GOES-18,GOES-18	2024-05-11 02:10	2024-05-11 01:00	10.9167	N/A	35.7863	2024-05-10 15:10	206.9193	2024-05-10 17:45
≥ 30	≥ 1	GOES-18,GOES-18	2024-05-11 02:00	2024-05-12 22:15	44.2500	N/A	27.2938	2024-05-11 07:20	39.2795	2024-05-11 09:10
≥ 50	≥ 1	GOES-18,GOES-18	2024-05-11 02:05	2024-05-12 11:20	33.2500	N/A	16.8888	2024-05-11 07:20	19.9175	2024-05-11 09:15
≥ 100	≥ 1	GOES-18,GOES-18	2024-05-11 02:10	2024-05-12 00:30	22.3333	N/A	7.1125	2024-05-11 06:55	7.7752	2024-05-11 07:15
≥ 10	≥ 10	GOES-18,GOES-18	2024-05-13 14:00	2024-05-16 14:55	72.9167	N/A	48.9364	2024-05-13 21:50	120.9137	2024-05-14 05:20
≥ 30	≥ 1	GOES-18,GOES-18	2024-05-13 11:45	2024-05-16 05:00	65.2500	N/A	8.6854	2024-05-13 19:05	15.2310	2024-05-14 00:05
≥ 50	≥ 1	GOES-18,GOES-18	2024-05-13 17:45	2024-05-14 07:20	13.5833	N/A	1.8397	2024-05-13 18:50	2.4630	2024-05-14 00:15

All Clear Contingency Tables

Values are given in the form X (+Y), where X is the all-time quantity, and Y is the quantity added from this week's results. X is inclusive of Y.

Model Category	Model Flavor	Hits	Misses	False Alarms	Correct Negatives	Forecasts	All-Time Report Link
GSU	All Clear	230 (+31)	140 (+30)	5779 (+153)	1809 (+44)	8437 (+321)	GSU All Clear
MAG4	LOS_FEr	488 (+60)	96 (+0)	12968 (+198)	9919 (+2)	24049 (+321)	MAG4 LOS_FEr
MAG4	LOS_r	436 (+60)	148 (+0)	10244 (+186)	12553 (+14)	24059 (+321)	MAG4 LOS_r
MAG4	SHARP	143 (+38)	259 (+15)	1761 (+121)	14508 (+45)	17065 (+271)	MAG4 SHARP
MAG4	SHARP_FE	165 (+38)	229 (+14)	2413 (+121)	13623 (+42)	16820 (+267)	MAG4 SHARP_FE
MAG4	SHARP_HMI	85 (+32)	327 (+22)	571 (+95)	15905 (+73)	17286 (+274)	MAG4 SHARP_HMI
MagPy	SHARP_HMI_CEA	1 (+0)	339 (+58)	93 (+23)	6611 (+188)	7481 (+329)	MagPy SHARP_HMI_CEA
SAWS-ASPECS	flare	69 (+22)	434 (+20)	233 (+2)	13306 (+22)	21157 (+99)	SAWS-ASPECS flare

Figure 2.6: Portion of MailSPHINX email; includes overview, event summary, and all-clear contingency table sections.

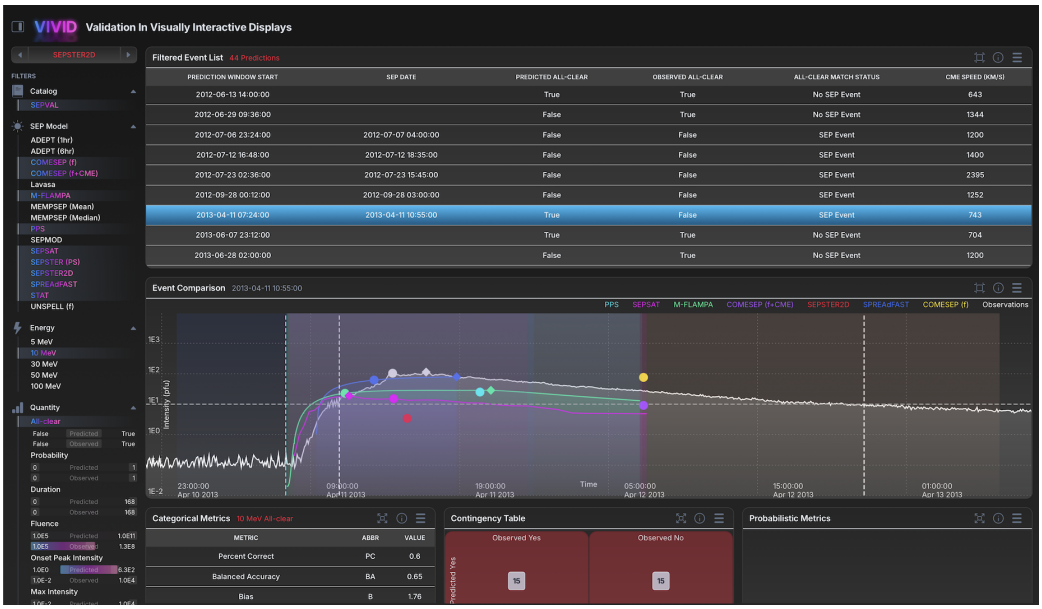


Figure 2.7: An example screenshot of VIVID showcasing filters in the sidebar, tables of results and metrics, and plots of models compared together. Note: not all filters, tables, and plots are shown – as some are in the scrollable area below the screenshot.

SPHINX is capable of validating a large set of predictions and observations. However, if any predictions need to be excluded from the validation study, SPHINX needs to be rerun or separate post-processing programs need to be applied in order to calculate the new metrics. VIVID solves this problem by taking the entire set of SPHINX results, then allowing the user to apply filters to create any subset of desired results. Metrics are then immediately recalculated for the subset. The filters within VIVID include catalog, SEP model, energy, quantity, date range, trigger, and match status. The catalog refers to the set of SPHINX validation results such as the one for SEPVAL or the SEP Scoreboard. The rest of the filters are based on what is provided in SPHINX. Any continuous quantity or trigger can be further filtered by minimum and maximum values. Similarly, categorical quantities such as all-clear can be filtered by True/False. Filtering within VIVID gives the ability to determine the performance of SEP models in any given environmental conditions.

During an SEP event, SRAG console operators monitor the SEP Scoreboards as new predictions are made. The operators need to verify how much trust they can put in these predictions since they contribute to the operator's recommendation for crew action. Metrics from SPHINX validation studies reflect model performance for a wide range of SEP events, but these metrics may slightly differ for the current, ongoing SEP event that the operators are monitoring. By using VIVID, SRAG console operators can select the models making predictions on the SEP Scoreboards, apply filters based on the current conditions of the ongoing SEP event, and verify their trust in the predictions based on the newly-calculated metrics. As an example, SEPSTER is a model that predicts the onset peak of an SEP event based on the

associated CME. If the CME was fast, wide, and occurred on the west limb of the Sun, then the console operator can use VIVID to filter out slow, narrow, and non-western CMEs. All metrics are then recalculated for the subset of SEPSTER predictions that only pertain to the current CME. This allows for a more precise representation of model performance for the ongoing SEP event.

An important approach to improving the SEP models is allowing the developers to explore the validation results in addition to the SRAG validation team. Since VIVID is an accessible web app, the model developers can explore independently. A commonly requested feature within VIVID is its custom plots where the user can plot any quantity or trigger on either axis. This is important for SEP models based on machine learning algorithms in order to determine if there is any internal correlation — which may be reflected in the metrics and therefore inaccurately represent the model performance.

VIVID offers a handful of additional features beyond exploring validation results. Users can download the subset of SPHINX validation results pertaining to the models and other filters selected. The intent of this tool is to allow the SRAG validation team, model developers, and others in the space weather community to use the results for independent studies. Another feature is the ability to download any of the plots that are displayed on VIVID so they can be used for publications and presentations. Since VIVID is an accessible web app and calculates metrics on-the-fly, it is a valuable tool for discussion-based working meetings. This was demonstrated by using VIVID in real-time during SEPVAL, our working group sessions at ISWAT, and Topical Discussion Meetings at ESWW.

To summarize the SPHINX Validation Framework:

- The collection of programs that comprise the SPHINX Validation Framework prepares satellite observations as ground truth, automatically matches forecasts to observations, calculates a wide variety of validation metrics, and generates reports and emails for end-users.

- VIVID is an stand-alone interactive web tool that allows users to investigate SPHINX results by applying filters, generating plots and recalculating metrics on the fly.

- SPHINX was developed to be robust to the unique challenges of real-time forecasting, including duplicate, incomplete, or corrupted forecasts, the ability to incrementally update validation results with new forecasts, and optimization to manage the hundreds of thousands of forecasts issued to the Scoreboards each month.

3 Validation Approach

3.1 Validation Overview

In this report, we present validation results for two separate efforts — SEPVAL and the SEP Scoreboards. The datasets and participating models are described in detail in Section 4. Each effort contributes distinct insights into model performance, with different benefits and drawbacks. Taken together, they provide a broad assessment of predictive skill and operational utility.

SEPVAL was carried out as a community challenge with the participation of SEP model developers who volunteered their time and effort. An approximately equal sample of SEP events and non-event periods were selected by SRAG, focusing on the types of SEP events and parent eruptions (flares and CMEs) that are relevant to SRAG operations. Nearly all of the selected periods were associated with strong flares and fast CMEs, except for three SEP events for which the flares were not visible from Earth because the source regions were beyond the Sun's western limb. Prior to the SEPVAL effort, many published validation studies focused on SEP events only without testing for quiet periods. A specific aim of SEPVAL was to include SEP events **and** time periods following solar eruptions that did not result in significant proton flux enhancements at Earth, referred to here as non-event periods, so that the validation would reflect correct predictions, misses, and false alarms.

The SEPVAL challenge encouraged developers of all types of models to participate. A few developers submitted science-quality predictions that had been optimized to produce the best results possible for each event (STAT and M-FLAMPA). These types of predictions answer the question "What is the best we can do with our current technology and understanding?". SEPVAL's primary interest was real-time performance and most model developers submitted forecasts in that category. Models were required to set free parameters to default values and produce forecasts without tuning or calibrating from event-to-event. Developers were asked to use flare and CME parameters provided by SEPVAL organizers or to use the real-time inputs their system was designed to use if it was already mature enough to run in real time. While this approach simulates a type of real-time workflow, the flare and CME parameters provided by SEPVAL could not represent what was truly available at the time of each event because catalogs do not sufficiently version control historical data. We instead focused on providing the best measurements possible. M2M reviewed each CME and produced high-quality 3-dimensional CME parameters with the same tools used to support SRAG operations. Models also had the option to use 2-dimensional plane-of-sky CME parameters published in the human-derived CDAW CME Catalog[4] or automatically-derived CACTus catalog[5]. In this way, SEPVAL allows cross-comparisons between models using their most appropriate workflows for a standardized set of challenge events.

The SEP Scoreboards, built to visualize forecasts in real time for use in SRAG console operations, represent a very different validation scenario. Models forecast-

[4] https://cdaw.gsfc.nasa.gov/CME_list/
[5] https://www.sidc.be/cactus/

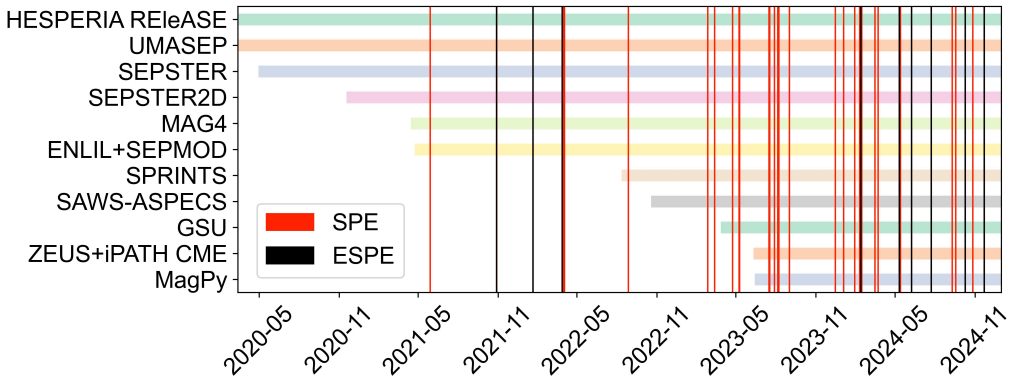


Figure 3.1: Timeline for models onboarded into the SEP Scoreboards. Red vertical lines indicate Solar Particle Event (SPE) where >10 MeV exceeds 10 pfu. Black vertical lines indicate Energetic Solar Particle Event (ESPE) where >100 MeV exceeds 1 pfu.

ing to the SEP Scoreboards use true real-time inputs as they become available from various satellites, ground-based observatories, SWPC forecasters, and M2M analysts. Real-time measurements are often of lower quality and prone to data gaps as opposed to the higher-quality, post-processed, back-filled archived data used in historical analyses. Models run regularly and automatically with humans in the loop only to provide necessary inputs (e.g. CME measurements). Here, model run-times play an important role in the ability to produce forecasts in a timely manner for SRAG operators. Forecast coverage represents the true climatological ratio of SEP events as rare deviations from typical quiet conditions. The forecasts aggregated on the SEP Scoreboards are the best representation of forecasting capabilities and performance of SEP models today.

The SEP Scoreboards are a valuable tool for SRAG, but they are also active platforms for research and development. Models within the Scoreboards are at varying levels of maturity. Many of the model developers have had to resolve bugs that resulted in corrupted forecast files, incorrect forecasted values, duplicate forecasts, missing forecasts or other problems. Throughout the lifetime of the Scoreboards, many models have upgraded versions to address problems or implement improvements, resulting in forecasts produced by a mix of versions for the same model. Furthermore, models have been onboarded on different dates and "encountered" different numbers of SEP events, resulting in incongruous forecast coverage. Figure 3.1 shows the time coverage for each model on the SEP Scoreboards and the SEP events (red and black vertical lines) that occurred in those time frames.

The characteristics of the source eruptions for the SEP event and non-event periods in SEPVAL and the SEP Scoreboards are shown in Figures 3.2 – 3.5. The events for the SEPVAL challenge were selected with the aim of creating similar distributions of flare and CME properties for the SEP event and non-event periods. The yellow (SEPVAL non-events) and blue (SEPVAL SEP events) in Figures 3.2

– 3.5 show that this goal was not quite accomplished but that there is significant overlap between the two distributions. The Scoreboard SEP events (green) and non-event flares and CMEs (red) show the true climatological distribution of eruptions for Solar Cycle 25. The Scoreboard non-event periods cover all C, M, and X flares and all CMEs with speeds ≥400 km/s and widths ≥10 degrees entered into DONKI from March 2020 to December 2024, representing the active lifetime of the Scoreboards.

Figures 3.2 – 3.5 show that the SEPVAL non-event flare and CME distributions are different from both the SEPVAL SEP event distributions and the Scoreboard non-event distributions. Model performance on the SEPVAL non-events can indicate whether a model has skill in discriminating between event and non-event periods, but is not necessarily representative of performance for a climatological sample of non-event flares and CMEs. The figures also highlight the true imbalance of reality — only 37 out of thousands of flares and CMEs produced threshold-crossing >10 MeV SEP events at Earth in Solar Cycle 25.

The SEP maximum flux distributions for SEPVAL and the SEP Scoreboards are shown in Figure 3.6. It is clear that SEPVAL contains larger SEP events than have occurred Solar Cycle 25 so far, particularly in the >100 MeV channel. For SEPVAL, 13/32 (41%) of >10 MeV events exceed 100 pfu and 5/32 (16%) exceed 1000 pfu, whereas only 13/37 (35%) exceed 100 pfu and 2/37 (5%) exceed 1000 pfu for the SEP Scoreboards. This is a potential source of bias that should be kept in mind when interpreting results from both datasets as models may have more success predicting larger events.

The two validation approaches of SEPVAL and the SEP Scoreboards provide different interpretations of model performance with different characteristics, outlined in Figure 3.7. We analyze these results to inform SRAG, ISEP, and the research community. In Section 4, we describe the SEPVAL and SEP Scoreboards models and datasets that contribute to the validation results. In Section 5, we discuss the ensemble median results for selected metrics. In Section 6, we focus on the models active in the SEP Scoreboards as they are currently being evaluated by SRAG for utility in space radiation operations. Here, performance is described in detail for the SEP Scoreboards and SEPVAL (when applicable) with interpretations and conclusions for each model. For similar models, cross-model comparisons are discussed in Section 7. Finally, state-of-the-art model performance is derived for historical (SEPVAL) and real-time (SEP Scoreboards) scenarios in Section 8.1.

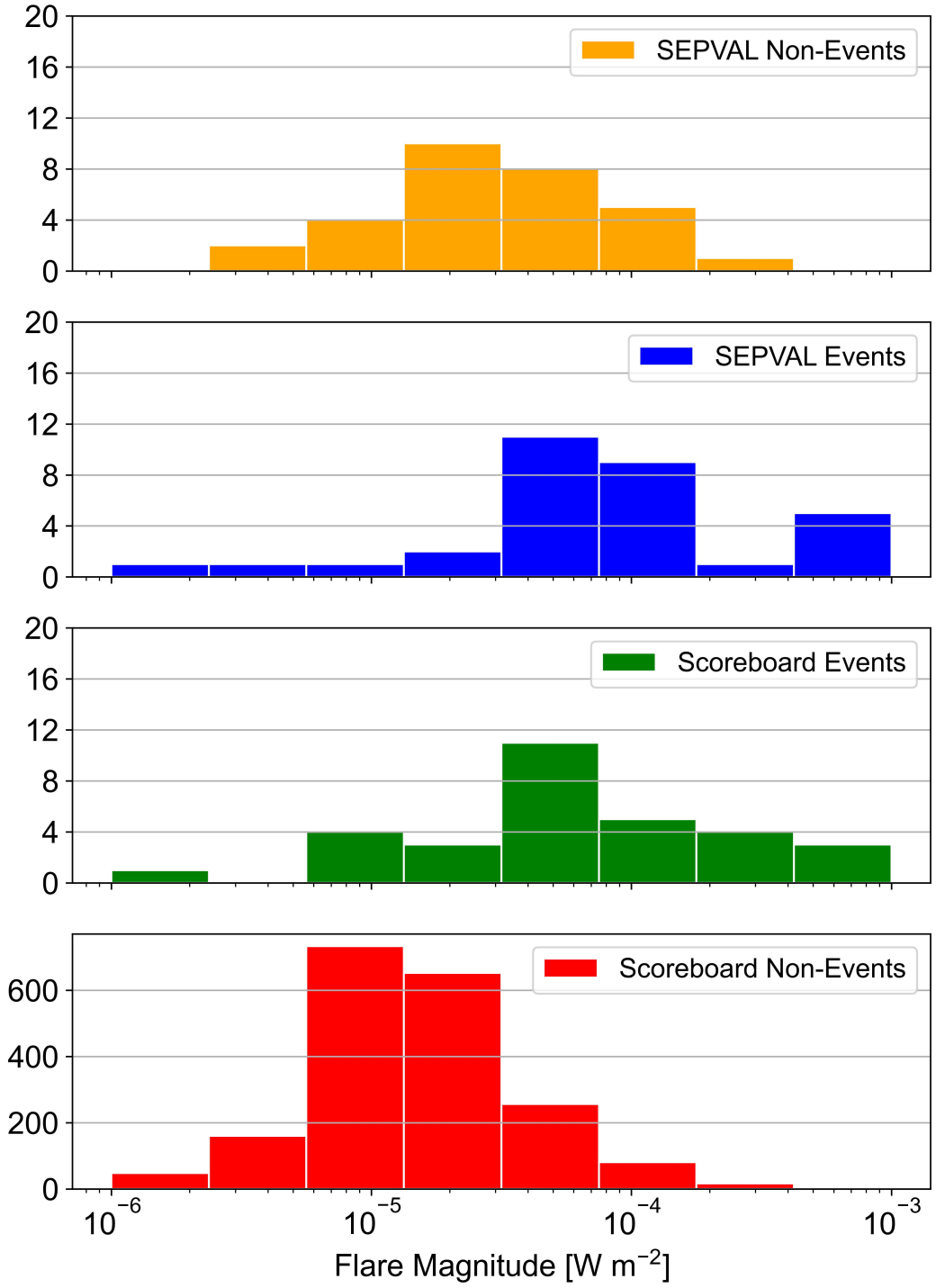


Figure 3.2: The distribution of flare magnitude for SEPVAL non-event periods (yellow), SEPVAL SEP events (blue), Scoreboard SEP events (green), and Scoreboard non-event flares.

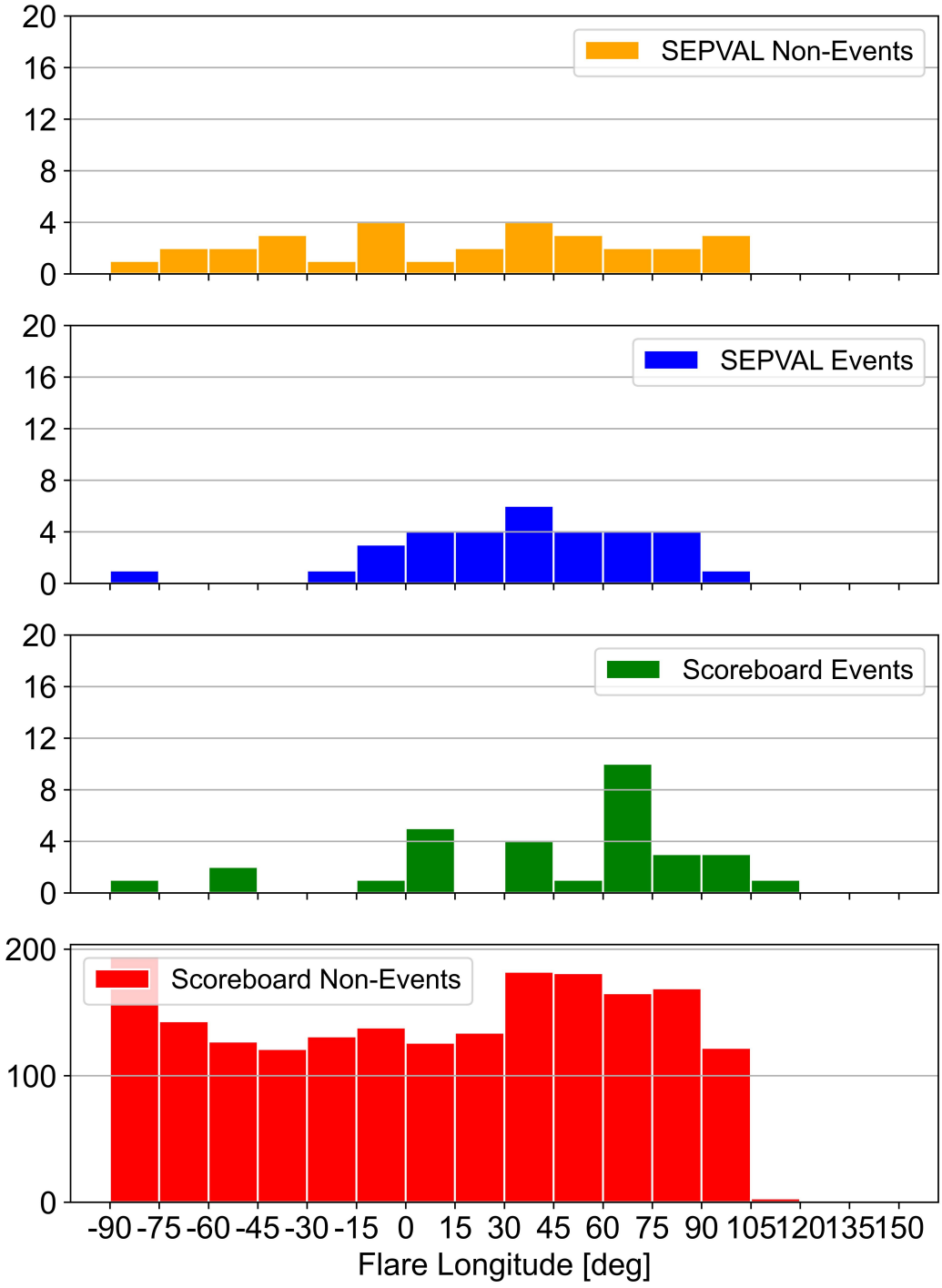


Figure 3.3: The distribution of flare longitude for SEPVAL non-event periods (yellow), SEPVAL SEP events (blue), Scoreboard SEP events (green), and Scoreboard non-event flares.

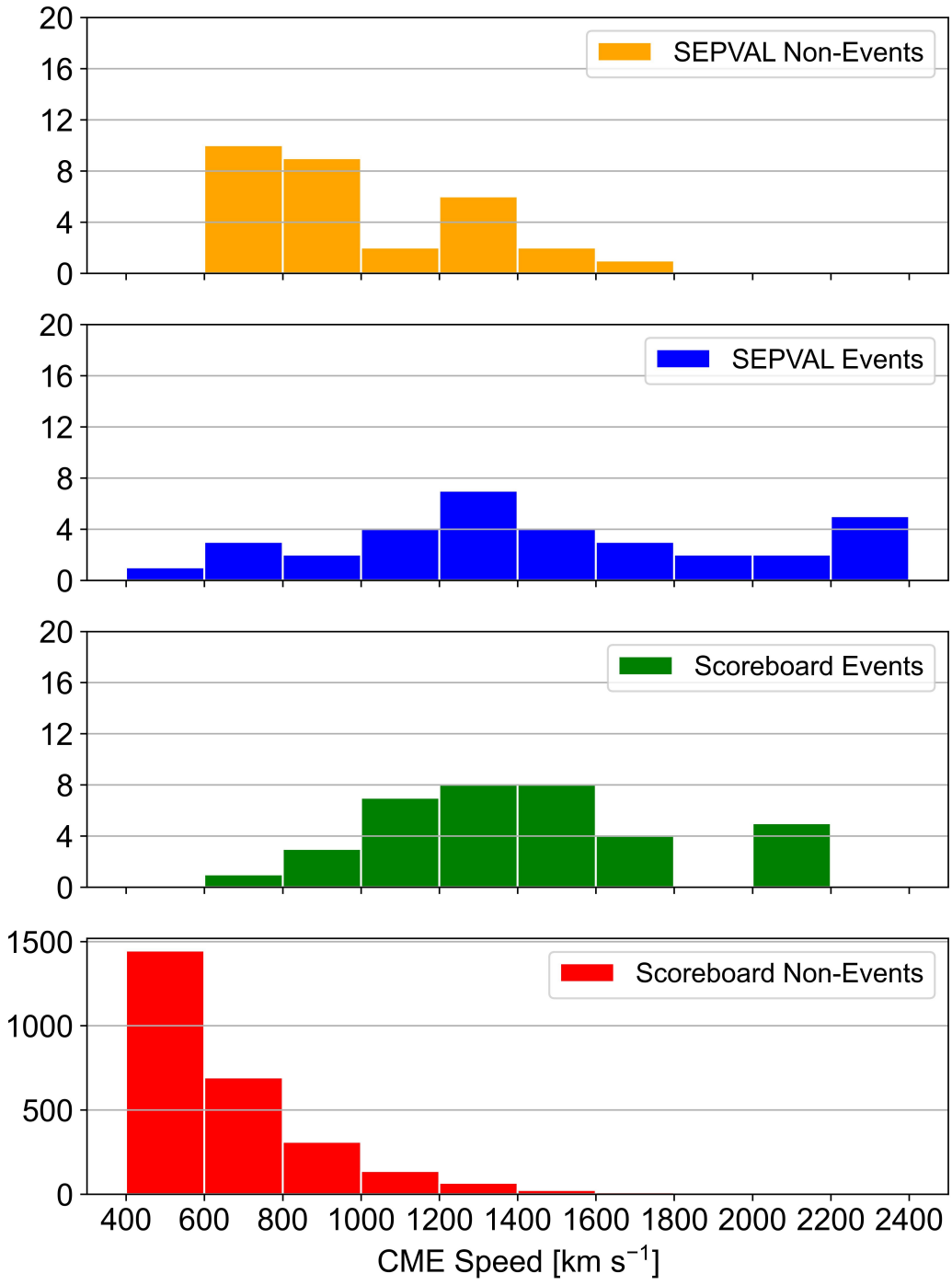


Figure 3.4: The distribution of CME speed for SEPVAL non-event periods (yellow), SEPVAL SEP events (blue), Scoreboard SEP events (green), and Scoreboard non-event CMEs with speed $\geq$400 km/s and width $\geq$10 degrees.

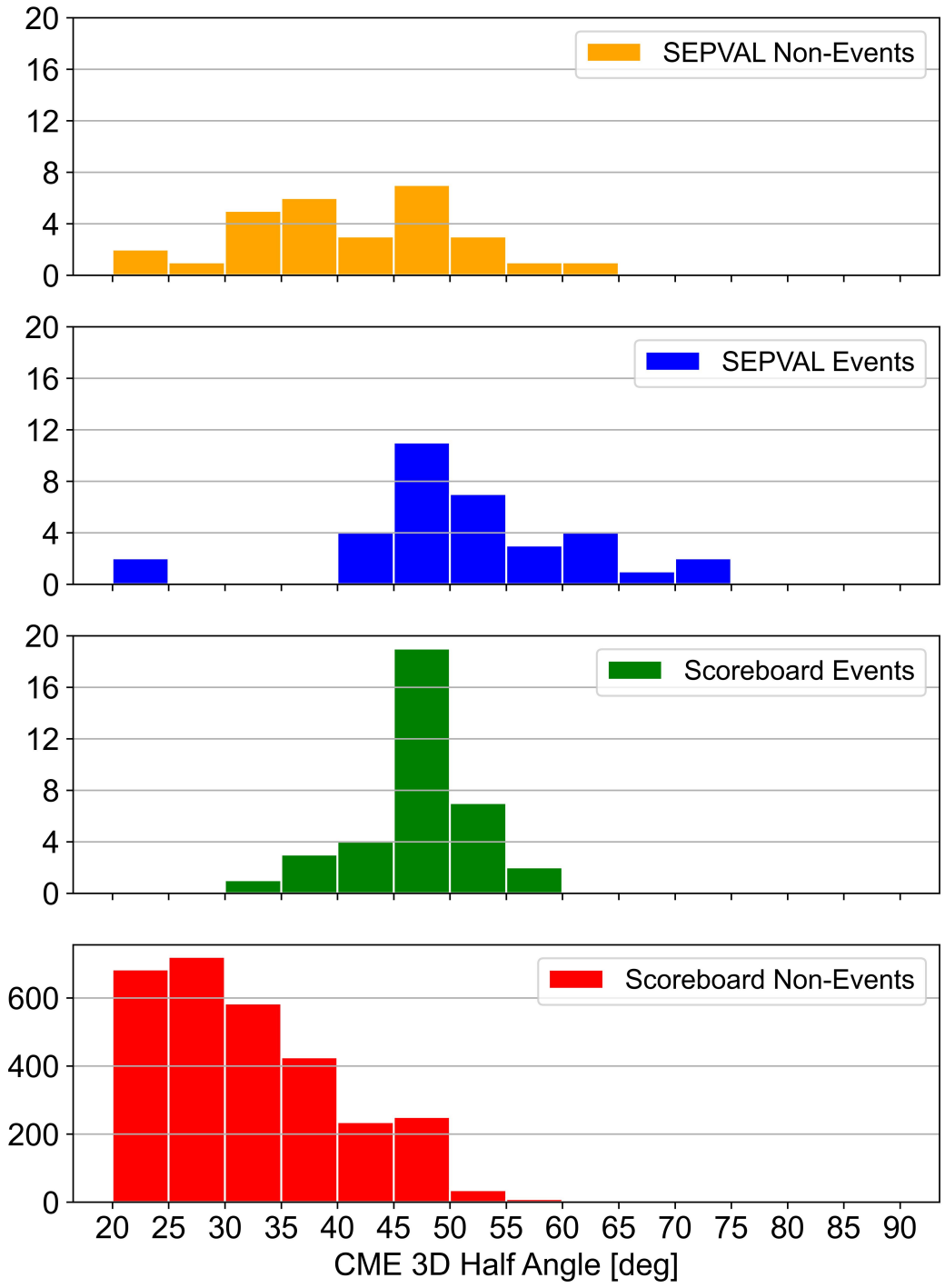


Figure 3.5: The distribution of CME 3D half angle from DONKI for SEPVAL non-event periods (yellow), SEPVAL SEP events (blue), Scoreboard SEP events (green), and Scoreboard non-event CMEs with speed ≥400 km/s and width ≥10 degrees.

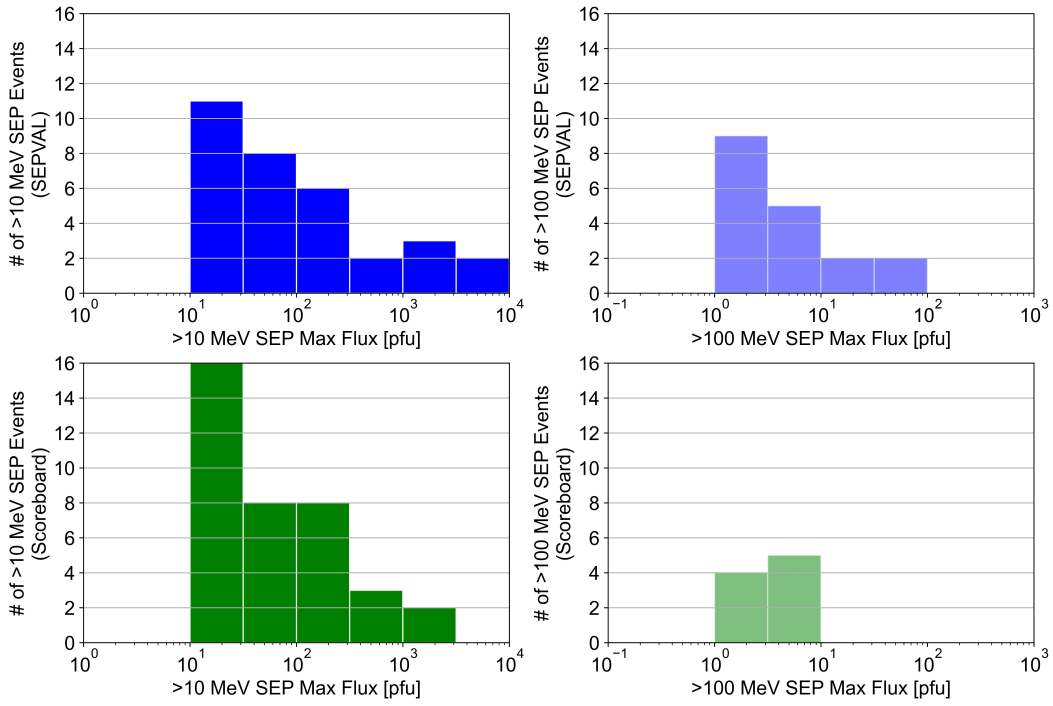


Figure 3.6: The distribution of maximum proton flux during the SEPVAL (blue) and SEP Scoreboard (green) SEP events for >10 MeV (left) and >100 MeV (right).

SEPVAL	Pro	Con
Inputs	High-quality, historical inputs	Not always available in real time
Dataset	Same for all models; allows cross-model comparisons	Balanced dataset not representative of true climatology
Dataset Bias	Stronger events relevant to space radiation operations	May not be representative of the general space weather environment, which can also vary according to solar cycle
Model Settings	Real-time-like workflow with default settings and no recalibration from event to event	Models may have included SEPVAL events in training/development
Model Output Files	SEPVAL team ensured forecast files were complete and contained intended forecasted values	No check of model robustness
Forecast Issue Time	Evaluated separately from performance, so realistic timing was not required for participating models to enable broader participation	Availability of forecasts in a timely manner could not be evaluated
Sensitivity Analysis	Enbles models to test various inputs or model settings for the same set of events to understand the effect on forecast skill	

SEP Scoreboards	Pro	Con
Inputs	True real-time inputs	Lower quality measurements with data latencies and gaps
Dataset	Reflects true space weather climatology during model forecasting period	Models cover different time periods; cross-model comparisons difficult
Dataset Bias	Representative of the space weather environment for SC 25	SEP events in SC 25 generally smaller than past cycles
Model Settings	True real-time workflow as would be used in operations	Model performance includes the effects of data availability, human-in-the-loop activities, and mixed versions
Model Output Files	Check of model robustness	Includes corrupted/incorrect/duplicate/missing forecasts
Forecast Issue Time	Representative of forecast availability to end-users	Impacted by effects beyond the model's control including data gaps and human availability
Sensitivity Analysis	Allows comparison between models with similar approaches	Detailed sensitivity analysis in real-time would require ensemble forecasts, which are not currently produced by any model

Figure 3.7: Pros and cons of the SEPVAL and SEP Scoreboard validation approaches.

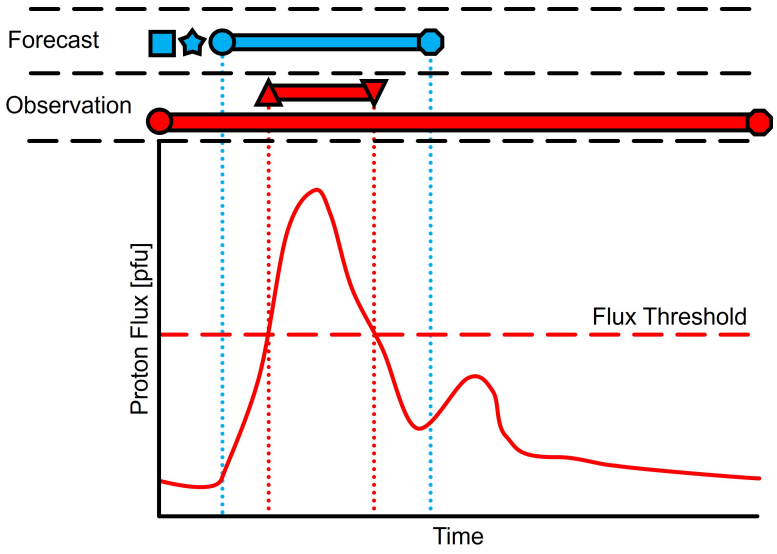
3.2 SPHINX Approach

The SPHINX Validation Framework was created to evaluate SEP model performance in a systematic, consistent, and fully reproducible manner. The logic within SPHINX is designed to validate forecasts with an emphasis on end-user needs. Within the automated logic, important choices are made that determine how observed values are paired with forecasted values for validation. When interpreting the validation results, it is worthwhile to be aware of the key parameters and caveats of SPHINX.

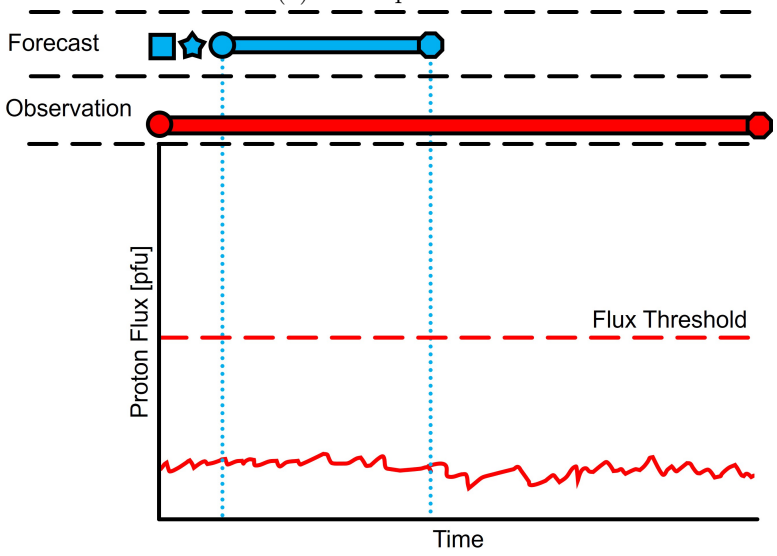
Figure 3.8 shows a simplified example of how SPHINX "evaluates" forecasts against observations. Some key components of a SPHINX evaluation are shown in Figure 3.8c. The forecast specifies a "prediction window" during which the forecast is valid — a time period within which the declarations made by the forecast (e.g., all clear, peak flux, probability of occurrence, etc.) are intended to apply (blue bar and blue horizontal lines). Given this prediction window, an observation window—a time period prepared by FetchSEP which encapsulates the prediction window and defines the range of available proton flux measurement data used for comparison with the forecast declarations—is also defined (red bar). With the prediction and observation windows defined, SPHINX is able to compare the forecast declarations against the observation data. Consider an SEP forecast model that attempts to predict whether an SEP event will or will not occur. Figure 3.8a shows an example of an observed SEP event with forecast and observation parameters overlaid. SEP event duration is defined as the duration over which the proton flux exceeds the flux threshold (red horizontal lines). In this example, the prediction window encapsulates the SEP event, the forecast was issued prior to the event onset, and the forecast was issued after the last piece of data necessary to issue the forecast was available. If this forecast had predicted that an SEP event would occur, then this forecast would be classified as a "Hit". If it failed to predict the onset of an SEP event, then this forecast would be classified as a "Miss". Contrast this with Figure 3.8b. In this example, the prediction window covers a period where the proton flux remains at background levels. If this forecast had incorrectly predicted that an SEP event would occur, then this forecast would be classified as a "False Alarm". If it had correctly predicted that no SEP event would occur in the prediction window, then this forecast would be classified as a "Correct Negative".

3.2.1 SPHINX Validation Philosophy

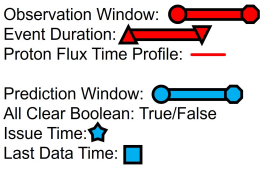
SPHINX asserts that a forecast is only considered a forecast if it uses measurements from the past or present, as indicated by the "last data time", to predict the future. Furthermore, a model must issue a forecast, specified by the "issue time", before the SEP event start (or peak, or end) has already occurred in order for the forecast to be used for real-time decision-making. SPHINX evaluates forecasts to quantify predictive skill and separately evaluates the availability of the forecasts in real time to quantify utility to the end user. SPHINX's strict requirements on timing may differ from validation studies published in the literature.



(a) Event period.



(b) Quiet period.



(c) Key.

Figure 3.8: Conceptual visualization of SPHINX "evaluating" a forecast against an observation.

3.2.2 Automation

To ensure reproducibility, and to handle the millions of forecasts issued to the SEP Scoreboards, SPHINX must be fully automated. As with any automated process, parameters set within the program are chosen to optimize correct outcomes, but occasionally result in incorrect outcomes. These errors introduce uncertainty in the final metrics. The most impactful choices and how they affect the final metrics are described here. **Substantial quality and assurance studies of SPHINX's performance were carried out by checking hundreds of forecasts associated with a sample of Scoreboard SEP events and checking the forecast and observation associations. Some errors were identified for a small percentage of forecasts, but for the vast majority, it was found that SPHINX assigned the correct observed values for comparison with the forecasts.** Nonetheless, it is important to understand the sources of uncertainty in the metrics, described below.

3.2.3 SPHINX Only Evaluates Forecasts Provided

SPHINX receives a list of forecasts as input. All forecasts undergo a series of quality control checks for completeness and timing and those forecasts that do not pass are removed from analysis. The remaining forecasts are then evaluated — they are paired with appropriate observed values, and then metrics are calculated to quantify model performance.

SPHINX is only aware of the forecasts in the input list, whether they indicate an event or not. If forecasts are not provided for a given SEP event, that event will not be included in the final metrics. If forecasts are not produced for quiet time periods, those quiet periods will not be included in the metrics. This means that, from SPHINX's perspective, each model in the SEP Scoreboards has its own individual climatology affected by any gaps in the input data or problems with the model pipeline and subsequent missing forecasts. Models use different inputs, are triggered by different phenomena, and have different cadences, so SPHINX does not have an automatic mechanism to know when a model *should have produced* a forecast. However, if a user would like to evaluate model performance for a specific set of time periods, the SPHINX framework includes post-analysis tools that can check a given list of SEP events or quiet periods and determine whether forecasts were issued on those dates and what their validation outcomes were.

Perhaps more critically, if predictions are triggered by a series of flares and CMEs that occur in quick succession followed by an SEP event, SPHINX expects that one of those forecasts is triggered by the parent eruption for the SEP. If for some reason, a forecast for the parent eruption was never issued, SPHINX does not know this and will associate an incorrect forecast with the observed SEP event. For example, consider three flares in a 24 hour period with one flare leading to an SEP event. If a model makes a prediction for only two of the flares, excluding the one that is known to be associated with the SEP event, SPHINX will take a forecast from an incorrect flare and associate it with the observed SEP event as long as the flare satisfies the matching criteria described in Section 3.2.7. When evaluating forecasts in real-time,

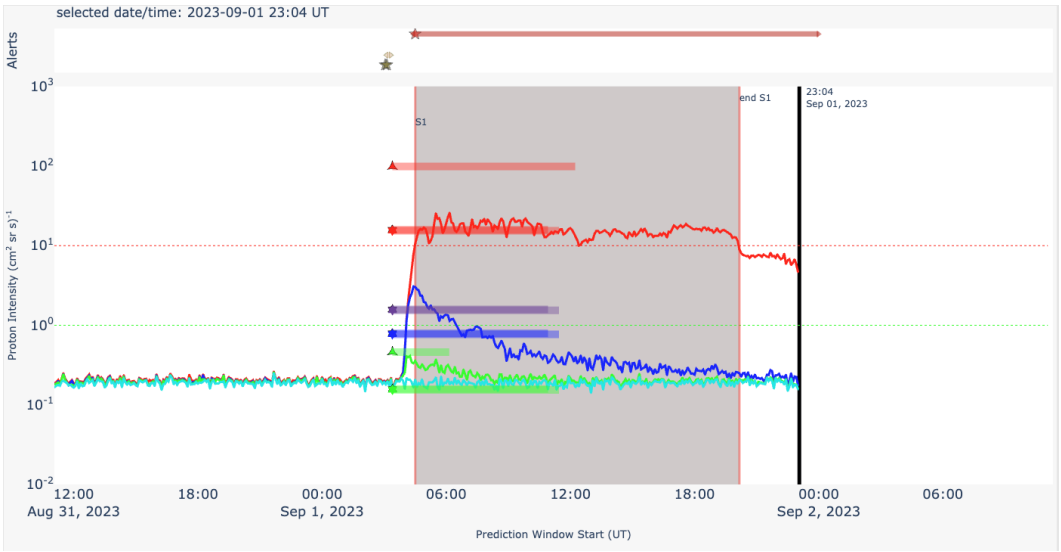


Figure 3.9: SEPSTER (star) and SEPSTER2D (triangle) forecasts with prediction windows (shaded bars) overlapping an observed SEP event (red vertical lines with shading in between).

SPHINX needs complete coverage to ensure correct associations. Work has already begun to resolve this problem by including the known flare and CME associations with observed SEP events in SPHINX's matching logic.

3.2.4 Last Data Time

Timing is important in SPHINX. Forecasts are required to include a "last data time" for all data ingested by the model for each forecast. To be considered a forecast in SPHINX, the last data time must occur before the forecasted phenomenon is observed. If a model predicts a threshold crossing time, Probability of Occurrence, or binary All Clear, then all data used as input into the model to make the forecast must be measured before the observed threshold crossing time. Likewise, if the prediction is for peak flux, then all input data must be measured prior to the observed peak. SPHINX will either discard a forecast with last data time after a phenomenon is observed or will assume the forecast is for a next event, depending on multiple factors. This may be different from validation published in the literature that may not apply such strict rules on timing.

3.2.5 Prediction Window

Each forecast specifies a "prediction window" that indicates the time period for which the forecast is valid, pictured in Figure 3.8. SPHINX will only compare forecasted values to observations within the prediction window. The choice of prediction window is thus very important. A model must choose an appropriate time frame that encompasses the observed phenomenon of interest.

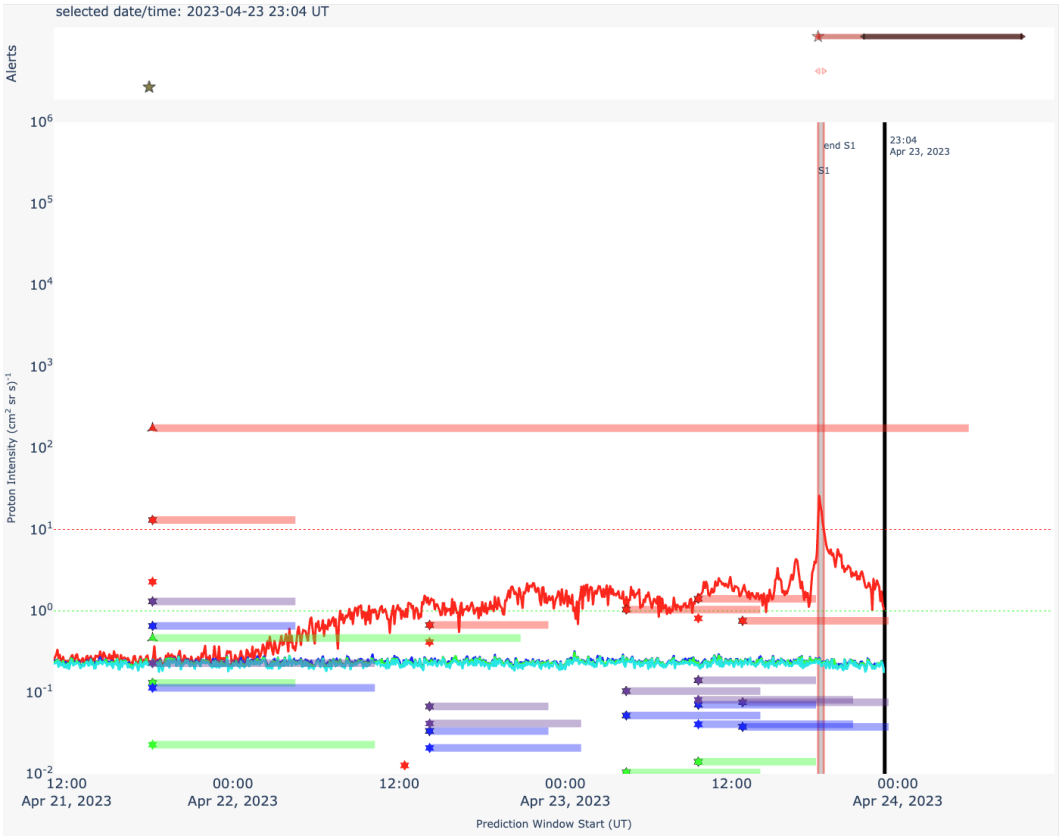


Figure 3.10: SEPSTER (star) and SEPSTER2D (triangle) forecasts with prediction windows (shaded bars). The SEPSTER prediction window for the leftmost forecast is too short to include the threshold crossing (red vertical lines with gray shading) for this very gradually rising SEP event even though the CME that triggered the forecast is responsible for the particle enhancement. The leftmost SEPSTER2D prediction window is long enough to capture the threshold crossing.

Figure 3.9 shows a case where Solar Energetic Particle STEReo (SEPSTER) and Solar Energetic Particle STEReo 2D (SEPSTER2D) forecasts are triggered by a CME and the forecast prediction window overlaps with the associated SEP event that follows. In this case, both SEPSTERs are assigned a hit. If a prediction window is too short, a forecast predicting an SEP would not get credit for a hit if the window did not extend to the observed SEP start time. Rather, such a forecast would be evaluated as a false alarm. Figure 3.10 shows SEPSTER and SEPSTER2D forecasts triggered by a CME (leftmost forecasts) and the very gradual SEP event that follows. The event finally crosses threshold nearly two days later likely due to an ESP caused by a passing CME at Earth. In this case, SEPSTER's prediction window is too short to overlap with the 10 pfu threshold crossing and it is assigned a false alarm. SEPSTER aims to forecast the onset peak, so in this case, this is the correct validation outcome since the onset peak does not exceed the threshold. Unfortunately, two SEPSTER forecasts near 12:00 on April 23, 2023 overlap with the observed SEP event, one of which will be incorrectly associated to the event and evaluated as a miss. The SEPSTER2D prediction window from the correct CME does overlap with the threshold crossing and is assigned a hit. SEPSTER2D aims to forecast the maximum flux, so this is the correct validation outcome. If a prediction window is too long, then correct forecasts for quiet periods would be evaluated as misses if a SEP event occurs inside the prediction window. The models evaluated in this report have generally applied appropriate prediction windows. In some cases, as demonstrated in Figure 3.10, there have been very gradually rising events that take a day or more to finally cross threshold. In those cases, the prediction windows for some models may not have been long enough to overlap with the observed SEP event even though the forecasts were triggered by the associated flare or CME.

3.2.6 Ongoing SEP Event

If a forecast's prediction window begins while particle fluxes are enhanced above threshold — specifically that the prediction window begins between the start and end time of an ongoing SEP event — SPHINX takes the approach that these forecasts cannot be evaluated. SPHINX assumes that a new forecast issued while an event is ongoing is forecasting for the next possible event. If no clear additional enhancement is seen in the particle fluxes, then it is not possible to determine whether a smaller, but possibly threshold-crossing enhancement occurred or the environment remained quiet. SPHINX chooses to remove these forecasts from the metrics and they count neither for nor against model performance.

3.2.7 Associating Eruption-triggered Forecasts with an SEP Event

Generally, if an SEP occurs inside of a forecast's prediction window, then SPHINX associates that forecast with the SEP event. If a model uses an eruption — a flare or CME — to trigger the forecast, additional logic is applied to determine whether that specific eruption could have reasonably caused the observed SEP event.

SPHINX requires flares and CMEs to occur within 24 hours prior to the proton threshold crossing to be associated with the event. For very gradually rising SEP

events or threshold crossings due to ESPs late in an event, this logic will not hold and forecast-to-observation matching errors may occur, as demonstrated in Figure 3.10. Three out of 37 events on the SEP Scoreboards fell into this category which may have resulted in incorrect associations or complete emissions of the events for some models. The events initiated by flares on 2023-04-21 (link to 2023-04-21 on the SEP Scoreboard) and 2024-09-14 (link to 2024-09-14 on the SEP Scoreboard) both cross threshold due to an ESP more than 48 hours after the initial particle enhancements began. The event on 2023-04-21 is shown in Figure 3.10 and matching outcomes for SEPSTER and SEPSTER2D were discussed above. The event on 2023-12-31 (link to 2023-12-31 on the SEP Scoreboard) rises very gradually, finally crossing threshold a few days later. Using SEPSTER as an example for 2023-12-31, no forecasts happened to overlap with the threshold crossing when it finally occurred, meaning that this particular SEP event does not appear in SEPSTER's validation results. The SEPSTER forecast triggered by the parent CME happened to predict that no threshold would be crossed. The observed onset peak remained below threshold inside of SEPSTER's prediction window and the forecast was appropriately assigned a result of correct negative since SEPSTER aims to predict the onset peak.

When forecasts triggered by different flares or CMEs are associated by SPHINX with the same SEP event, additional logic is applied to identify the most appropriate source eruption. Strong eruptions of western M & X-class flares and CMEs with speeds $\geq$600 km/s and half widths $\geq$35 degrees are given preference. If there are multiple, the strong eruption closest in time to the SEP threshold crossing is taken as the forecast to associate with the event. If two strong flares and CMEs occur within 1.5 hours or less of each other (e.g., 2012-03-07), SPHINX will associate both with the following SEP event because it is not obvious which eruption is the source. If none of the forecasts are produced by strong eruptions, then the one closest in time to the threshold crossing is selected.

3.3 Description of Validation Metrics

There are dozens of metrics calculated within the SPHINX framework, which are used to determine performance across the variety of different quantities that are predicted for SEP events. A full description of all of these metrics is in Appendix A, but this section will focus on a subset of the metrics that are used in the validation described within this report. There are plans in place to add uncertainties for each of the metrics to allow users to better interpret whether differences in metric values are significant when making cross-model comparisons or evaluating improvements to model performance.

3.3.1 All Clear

In this report, All Clear is interpreted as a binary prediction for SEP proton flux to cross a threshold, answering the operational question: "Will an SEP event occur/not occur?" The metrics used for All Clear are a combination of metrics that are commonly used within the SEP modeling community and a number that have been determined by SRAG to be useful in addressing our operational questions. Before

	Observed		
	Yes	No	Sum
Pred. Yes	h	f	$h+f$
Pred. No	m	c	$m+c$
Sum	$h+m$	$f+c$	N

Table 3.1: Example contingency table.

describing our selected metrics, many are affected by the balance of the dataset, which is the observed frequency of "yes" events and "no" events. Balanced datasets have a 1:1 ratio of "yes" to "no," whereas SEP event climatology is much lower. In cases like the SEP Scoreboard, there can be hundreds of "no" forecasts (quiet periods) for each "yes" forecast (SEP events), which can affect how useful some metrics can be. A full description of this problem is included in Section 3.4.

All of the metrics for the All Clear quantity are defined in Collaboration for Australian Weather and Climate Research (2015) and the references therein[6]. All Clear metrics are defined from the components of the contingency table: hits (h), misses (m), false alarms (f), correct negatives (c), and total forecasts (N). The below definitions use the description of "yes" events to be SEP events for which the proton flux crossed threshold, and "no" events to be periods when proton flux remained at background or below threshold.

All Clear metrics are derived from the components of the contingency table (Table 3.1),

- h (hits) represents the number of events forecasted to occur and actually occurred,
- m (misses) represents the number of events forecasted to *not* occur and actually occurred,
- f (false alarms) represents the number of events forecasted to occur and *did not* actually occur,
- c (correct negatives) represents the number of events forecasted to *not* occur and *did not* actually occur, and
- N is the total number of forecasts.

The all clear metrics that SPHINX calculates are derived from these integers.

The first set of metrics are ratios, which describe some fraction of events, meaning they all have ranges from 0 to 1 and the perfect score can be either depending on if that fraction needs to be small or large. **Percent Correct** is the fraction of forecasts that are correctly forecast (events are hit and non-events are correctly negative):

$$\text{Percent Correct} = \frac{h+c}{N}, \tag{3.1}$$

[6] https://www.cawcr.gov.au/projects/verification/#Methods_for_probabilistic_forecasts

also known as accuracy and has a perfect score of 1. **Hit Rate** is the fraction of "yes" events that are correctly forecast as "yes", in other words, SEP events that are hits:

$$\text{Hit Rate} = \frac{h}{h+m}, \tag{3.2}$$

is also commonly called recall, sensitivity, or Probability of Detection (POD). This metric is not affected by the imbalance of the dataset, and is important since it determines whether a model is responding to the environmental conditions that produce SEP events. A high hit rate (close to 1) is desired. **False Alarm Rate** is the fraction of "no" events that were incorrectly forecasted as "yes":

$$\text{False Alarm Rate} = \frac{f}{f+c}, \tag{3.3}$$

also known as the False Positive Rate. Typically used alongside the false alarm rate is the **False Alarm Ratio (FAR)** which is the fraction of "yes" forecasts that are false alarms:

$$\text{False Alarm Ratio (FAR)} = \frac{f}{h+f}. \tag{3.4}$$

In highly unbalanced datasets like the SEP Scoreboards, the False Alarm Rate is dominated by the large number of negative event periods. FAR is more sensitive to false alarms due to the very small number number of positive observed events—even a small False Alarm Rate can result in a large FAR in a highly imbalanced dataset. For both of these false alarm metrics lower scores are better with 0 being the best. The last of these fractional scores is the **Threat Score** which answers the question: "How well did the forecast 'yes' events correspond to observed 'yes' events?"

$$\text{Threat Score} = \frac{h}{h+f+m}. \tag{3.5}$$

is also commonly called the Critical Success Index. Threat score is another accuracy score, but one that does not include correct negatives while still being sensitive to hits, false alarms and misses, the latter two penalize the metric. In this way, the Threat Score sidesteps the problem of being swamped by correct negatives for highly imbalanced climatologies. **Bias** indicates a model's tendency to overforecast (false alarms) or underforecast (misses) events.

$$\text{Bias} = \frac{h+f}{h+m}. \tag{3.6}$$

Bias has a range from 0 to infinity, and a perfect score is 1. Values of bias less than 1 mean a model has a tendency toward misses, whereas values greater than 1 indicate the model has a tendency towards false alarms.

The next type of metric is a skill score, which is defined as a measure that looks at the relative improvement of the forecast over some reference forecast.[7]. Skill scores often range from -1 to 1, with 1 being a perfect score, 0 represents no skill, and

[7] https://www.cawcr.gov.au/projects/verification/#Skill_score

−1 being perfectly negative skill (always a "yes" forecast for "no" events, and vice versa). The reference forecast for **Heidke Skill Score (HSS)** is that of random chance (sometimes this is replaced with a climatological forecast):

$$\text{HSS} = \frac{2(hc - fm)}{(h+m)(m+c) + (h+f)(f+c)}. \tag{3.7}$$

HSS is very sensitive to the class imbalance of SEP events, meaning that its numerical value cannot be compared across datasets with different imbalances. HSS is a good indicator of the type of skill important for SRAG, showing sensitivity to the number of hits and false alarms even when the number of correct negatives is very large. **True Skill Statistic (TSS)** is a measure of discrimination used to answer "how well did the forecast separate the 'yes' events from 'no' events?"

$$\text{TSS} = \text{Hit Rate} - \text{False Alarm Rate} = \frac{h}{h+m} - \frac{f}{f+c}. \tag{3.8}$$

and is also referred to as Hanssen and Kuipers discriminant or Pierce's skill score. Notably, this metric is not affected by class imbalance in the sense that it will give the same score for models with the same Hit Rates and False Alarm Rates, regardless of class imbalance. However, a False Alarm Rate of, e.g., 25% may be acceptable in a balanced data set, but far too large for an imbalanced data set due to the very large number of false alarms. TSS has disadvantages in situations where events are rare Doswell III et al. (1990). This score is widely reported in the research community, but is not a sensitive indicator of the aspects of model performance that are important to SRAG. There are many other All Clear metrics that are calculated by SPHINX, which are listed in Appendix A

In light of operational needs, SRAG proposes a new skill score called False Alarm Event Ratio (FAER), pronounced "fear", where:

$$FAER = \frac{f}{h+m} \tag{3.9}$$

This ratio represents the number of false alarms compared to the number of observed events. FAER ranges from 0 to infinity and, ideally, it should be less than 1 and close to zero. For values greater than 1, it represents the excess factor of false alarms to the number of observed SEP events. For example, FAER = 30 means there are 30 times more false alarms than observed SEP events in the validation dataset. This metric has a simple intuitive utility for communicating to operators as models with high FAER are impossible for operators to trust.

3.3.2 Probability

The next set of metrics address models that provide forecasts of the probability that an SEP event will occur. Many of these models also provide an All Clear binary forecast by evaluating whether the forecasted probability is above or below an internally applied probability threshold. The first probability metric is the **Brier**

Score, which is equivalent to a mean squared error and measures the magnitude of the probability errors.

$$\text{Brier Score} = \frac{1}{N}\sum^{N}(P_{predicted} - P_{observed})^2, \tag{3.10}$$

where $P_{predicted}$ is the predicted probability, $P_{observed}$ is the observed probability (0 for "no" events and 1 "yes" events) and N is the total number of prediction and observation pairs. A perfect score for Brier Score is 0, and due to this metric being sensitive to the class imbalance of SEP events it is easy for models with no skill to have good scores. The **Brier Skill Score** uses the previous Brier Score to develop a skill score to describe the relative skill of the model compared to a reference forecast:

$$\text{Brier Skill Score (BSS)} = 1 - \frac{\text{Brier Score}}{\text{Brier Score}_{reference}}. \tag{3.11}$$

Unlike other skill scores mentioned previously, the range of the metric is from negative infinity to 1, with a perfect score of 1, and a score of 0 showing no skill compared to the reference forecast. In this validation report, the reference forecast used in the Brier Skill Score is 3.3% from Bain et al. (2021), representing the average likelihood of an SEP event for any given day in Solar Cycle 24. This is an appropriate climatological comparison for pre-eruptive models that issue forecasts on a regular cadence, but it is not an appropriate climatological reference for models triggered by flares and CMEs. Using this percentage affects our validation in two ways: 1) for the balanced dataset of the SEPVAL challenge, this climatology is much lower than the frequency of "yes" events and 2) since this percentage is from Solar Cycle 24, it does not represent the climatology of Solar Cycle 25 which cannot be calculated until the solar cycle has completed.

Used frequently with probabilistic models is a **Receiver Operator Characteristic (ROC)** curve. An ROC is a plot that displays how well a model can differentiate between "yes" and "no" events for probabilities, and an example is shown in Figure 3.11. On the x-axis is the False Alarm Rate and the y-axis is the Hit Rate, such that each point represents the False Alarm Rate and Hit Rate of the model when the internal probability threshold is varied to determine the binary All Clear value. A perfect forecast on this plot would have a vertical line at False Alarm Rate of 0 up to Hit Rate 1 ((0,0) to (0,1))and a horizontal line from Hit Rate 1 out to False Alarm Rate 1 ((0,1) to (1,1)). A model predicting a purely random guess would have a slope of 1/2 and would bisect the area of the plot evenly. Models that have more skill than random chance would produce a curve above the random guess line, and models with less skill would be below. Integrating under the ROC curve produces the **Area Under the Curve (AUC)**, which is commonly used as a complimentary metric. The random guess line has a value of 1/2, the perfect forecast has a value of 1, and the minimum value is 0. Models demonstrate skill if AUC is greater than 0.5 and it is desirable if the curve tends towards the upper left of the plot such that there is a probability threshold for which the False Alarm Rate remains low but the Hit Rate is high.

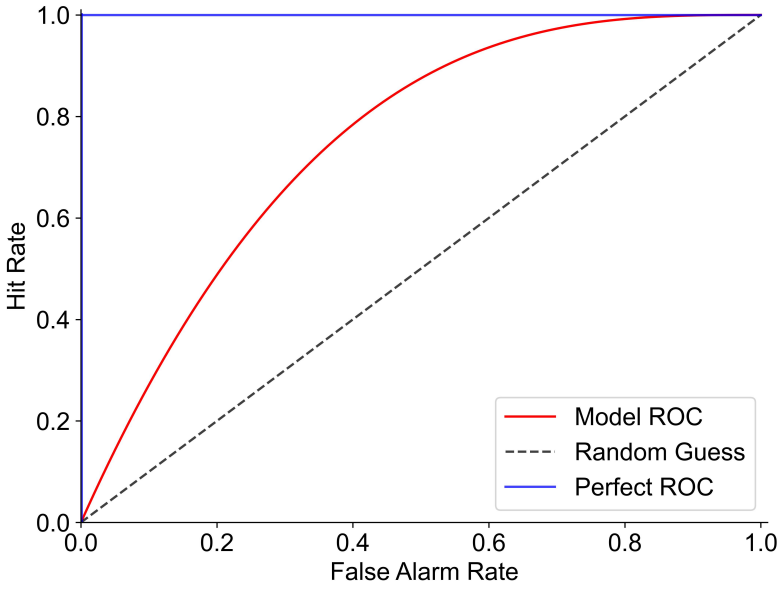


Figure 3.11: Example ROC Curve demonstrating a perfect forecast, random guess, and a hypothetical model curve.

Another plot that is used with probabilistic models is a **reliability** diagram shown in Figure 3.12. This plot displays of the frequency of observed events for each predicted probability. On the x-axis is the predicted probability value and on the left y-axis is the observed frequency of events that occur within the prediction windows when the prediction is at that probability value. A perfect probability model would generate points along the diagonal line, where the predicted probability value corresponds to the same frequency of observed events (when the predicted probability is 30% and 30% of the time there is an observed event in the same time frame). Displayed in the background is a histogram of the counts for each probability bin, with the number of counts displayed on the right y-axis. The histogram is used in tandem with the line plot to understand the frequency at which the model predicts certain probability values and whether there are sufficient statistics to calculate reliable observed frequencies.

3.3.3 Peak Flux

The next question that SRAG is asked in operations is: "How big will an SEP event be?" The answer to this question could be informed via forecast models that give a peak intensity prediction. In this report, we consider two types of peaks — the onset peak, which is the peak typically associated with the initial rise of an event (see Section 2.1), and/or the maximum peak, which is the peak flux achieved over the full duration of the event. For the observed onset peak, as well as the models that provide time profiles, we use an automated algorithm to find the onset peak as part of the FetchSEP module. The maximum peak flux during an SEP event is simply

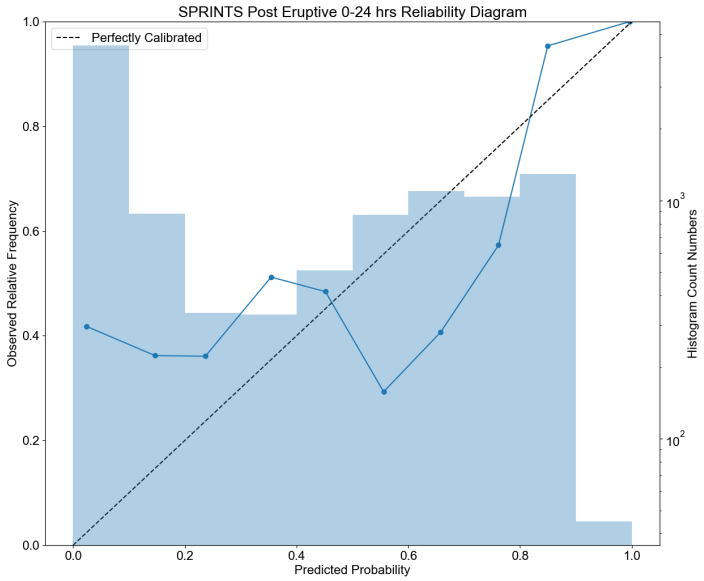


Figure 3.12: Example reliability plot.

the maximum value taken from the time profile of the respective energy channel, and for lower energies (i.e. >10 MeV for this report) this can be associated with the arrival of an Interplanetary Coronal Mass Ejection (ICME) at Earth called an ESP phase. Since these values of peak flux are given as flux values (with units of Particle Flux Unit (cm^{-2} s^{-1} sr^{-1}) (pfu)) the metrics used to determine skill are typically error metrics. Since peak values for a given energy channel can span multiple orders of magnitude across events (i.e. a small SPE event may peak near ∼10 pfu but a large one can be >1000 pfu), errors in linear space would be dominated by the larger events. To accommodate for this, we use **Log Error** to determine bias of the peak flux forecasts and **Absolute Log Error** to determine accuracy. We use the log-space errors since they better measure forecast order of magnitude error and are normalized across different orders of magnitude. Additionally, we look at histogram distributions of the errors as well as scatter plots to assess model performance in this quantity. Another measure used is the ratio or percent of forecasts that are within an a factor of 10 or 2 of the observed peak value. Many models forecast for only one of the peaks, but the peak intensity field in the CCMC JSON files is used ambiguously, so we provide metrics comparing the predicted value to both the onset peak and maximum flux, allowing the end-user to choose the appropriate one after the metrics have been calculated.

Related to peak intensity predictions is the third question SRAG wants to answer operationally, which is: "Do models capture the event-to-event variability of SEP events?" This question essentially asks if models know what causes an SEP event to be small (barely cross threshold) versus what causes an event to be large.

This question may be addressed by looking at correlation coefficients and linear regression, when appropriate. The **Pearson correlation coefficient** measures the linear correlation between two variables and is given by

$$\rho_{X,Y} = \frac{\text{cov}(X,Y)}{\sigma_X \sigma_Y} \tag{3.12}$$

where $\rho_{X,Y}$ is the correlation coefficient, $\text{cov}(X,Y)$ is the covariance of the two variables, and σ_i is the standard deviation of each variable (Weisstein, 2025). It is desirable for predicted and observed flux values to have a linear relationship in log space, thus the Pearson correlation coefficient is useful to indicate the level of linear correlation between the two. It should be noted that it is susceptible to outliers. The **Spearman correlation coefficient** (Spearman, 1904) measures the monotonic relationship between two variables without requiring any specific form of the relationship between them and is more robust to outliers. It is calculated by organizing the variables in rank order (smallest values are given the lowest rank and highest into high rank) then calculating the Pearson correlation coefficient using the ranks of the forecast and observation pairs. If the two data sets share the same rank order, then the Spearman correlation coefficient will have a high value of 1. If the rank orders are not correlated, then Spearman will be 0. If variables have opposite ranked order, then Spearman will be -1. Spearman will have a high value if the forecasts and observations have a monotonic relationship, but it will not indicate what type of relationship exists between them. The Pearson correlation coefficient then provides complimentary information about the linearity of the correlation. Note that correlations calculated from very sparse data hold very little meaning, so they should not be used unless there are enough data points. The literature reports minimum sample sizes of $n = 25$ or more (Bonnett and Wright, 2000), however our analyses are limited by small numbers of SEP events; we proceed with the data available. The reader should understand that the correlation values will become more reliable as statistics accrue. In the case that there are enough forecasts and observations and they show some level of correlation, **linear regression** in log space may provide useful information about the relationship between them. In particular, the value of the slope (best = 1) indicates whether the model appropriately captures the relative differences from event to event. It should be noted that outliers can significantly impact the linear regression line. Correlation coefficients and line fits should be interpreted with care, taking into account the sample size and presence of outliers. This is why it is best to use the correlation coefficient in tandem with the error metrics described above to analyze flux predictions.

3.3.4 Advance Warning Time

The last question SRAG wants to answer is: "Do models provide advance warning?" Forecasts hold no operational value if they are issued after the event has already unfolded in real time. To determine forecast availability, we calculate the Advance Warning Time (also sometimes called the lead time or forecast horizon) by comparing the forecast issue time (when the forecast is made available to the user) to the

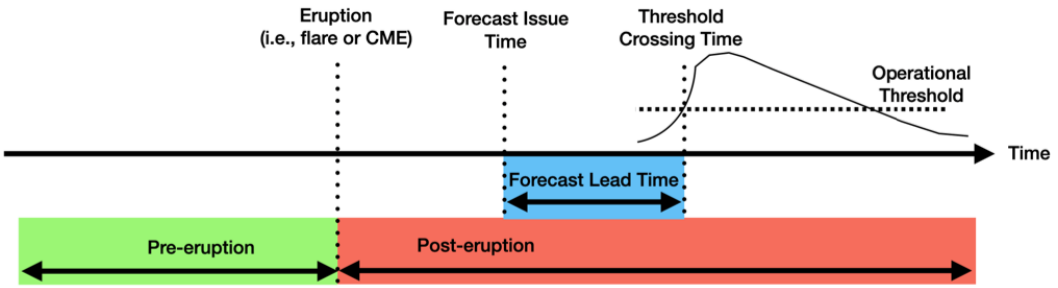


Figure 3.13: The AWT (or lead time) is the difference between the forecast issue time and SEP threshold crossing time. (Figure credit: Philip Quinn)

SEP event start time when the fluxes first cross threshold. Some models predict onset peak or maximum flux, so AWT with respect to the observed peaks is also calculated for those models. Even if a model cannot produce a forecast prior to the start of an event, a peak prediction is still useful before the observed peak occurs.

Figure 3.13 shows a schematic of the timings involved in a model forecast and observed SEP event. The AWT (labeled Forecast Lead Time in the figure) will be positive if the forecast is issued prior to the observed threshold crossing time and negative if it is issued after the threshold has already been crossed. Likewise for AWT with respect to peak values — if a forecast is issued prior to (after) the peak, AWT will be positive (negative).

There are many factors that contribute to the AWT of a model, from the timings of the event (e.g., the gradual or prompt nature of the particle rise after the solar eruption) to the real-time availability of a model's input data source, human-in-the-loop activities, and computation time. Due to this, AWT reflects the entire forecast pipeline.

As one last note on metrics, model performance in one quantity is not indicative of model skill or usefulness in operations, rather a comprehensive analysis across all of a model's forecasted quantities using a variety of metrics determines skill.

3.4 Impact of Imbalance on Metrics

Some All Clear and probability metrics are affected by the level of imbalance in the validation data set. The two validation cases presented here, SEPVAL and the SEP Scoreboards, have very different class imbalances. SEPVAL is a nearly balanced data set with ≈1:1 SEP events to non-event periods. The SEP Scoreboards, depending on the type of model and time period covered, have class imbalances closer to 1:28 for flare or CME-triggered models and even higher imbalances for pre-eruptive models like MAG4 and MagPy.

Ahmadzadeh et al. (2023) look closely at the behavior of various metrics with respect to True Positive Rate (TPR) (also known as Probability of Detection (POD) or Hit Rate (HR)) and True Negative Rate (TNR) (also known as Probability of Correct Negatives (POCN)). They show that, given the same TPR and TNR, some binary categorization (e.g., All Clear) metrics, such as HSS, result in different nu-

meric values for data sets with different class imbalances while other metrics, e.g. TSS, are invariant to class imbalance in this space. While this in and of itself is not necessarily a drawback for a given metric, it does indicate that HSS numeric values are not directly comparable between the SEPVAL and SEP Scoreboards. Furthermore, Ahmadzadeh et al. (2023) showed that the metric space for some metrics is distorted by class imbalance (e.g. Figure 5 in their paper) such that a small change in TPR or TNR might cause a large change in HSS or vice versa, depending on where model performance is within the space. This means that cross-model comparisons need to be done carefully in the context of class imbalance to understand whether a difference in scores between two models is a significant difference in performance according to end-user needs.

To demonstrate the effect of class imbalance on metrics, let us assume two contingency tables with TPR = 0.75 and TNR = 0.75, but different class imbalances. Using 32 SEP and 32 non-events (similar to SEPVAL), this gives the contingency table in Table 3.2 (left). If we assume a daily forecast for the past 4 years (1460 days) with 32 SEP events (similar to the SEP Scoreboards), this produces the contingency table in Table 3.2 (right). The resulting metrics for both cases are shown in Table 3.3. For the balanced contingency table, HSS=0.50, Threat Score=0.60, and TSS=0.50. For the imbalanced contingency table, HSS=0.079, Threat Score=0.062, and TSS=0.50. **This example highlights that the numerical values of HSS, Threat Score, and other metrics sensitive to imbalance should not be compared across SEPVAL and the SEP Scoreboards and one must be careful when comparing published results in the literature.**

Balanced ≈1:1

	Observed		
	Yes	No	Sum
Pred. Yes	24	8	32
Pred. No	8	24	32
Sum	32	32	64

Imbalanced ≈1:40

	Observed		
	Yes	No	Sum
Pred. Yes	24	357	381
Pred. No	8	1071	1079
Sum	32	1428	1460

Table 3.2: Balanced (left) and imbalanced (right) example contingency tables with Hit Rate (TNR) = 0.75 and TPR = 0.75. The numbers in the contingency tables are selected to be similar to the SEPVAL periods and the number of days models have been forecasting on the SEP Scoreboards.

The HSS and Threat Score are sensitive to the relative *number* of hits compared to false alarms. For SRAG's purposes, this is a desirable trait, since a large number of false alarms will erode operator trust in a model. The TSS reflects the *ratios* of correctly forecasted SEP events and correctly forecasted negative events, but may not fully reflect SRAG's needs as the relative numbers of SEP events and clear periods are not captured by this metric. For a model to achieve HSS = 0.50 for the same imbalanced data set, it would need a False Alarm Rate of 0.026. Scores for such a model are shown in the third column of Table 3.3. In a highly imbalanced dataset, it is important from SRAG's perspective for models to reduce the false alarm rate

All Clear Scores	Balanced ≈1:1	Imbalanced ≈1:40	Imbalanced High Skill
Percent Correct	0.75	0.75	0.97
Hit Rate	0.75	0.75	0.75
False Alarm Rate	0.25	0.25	0.026
False Alarm Ratio	0.25	0.94	0.61
Bias	1.0	11.9	1.9
Threat Score	0.60	0.062	0.35
HSS	0.50	0.079	0.50
TSS	0.50	0.50	0.72

Table 3.3: All Clear metrics for balanced and imbalanced datasets with Hit Rate (TPR) = 75% and TNR = 75% (False Alarm Rate = 25%) with the example contingency tables in Table 3.2. A third column shows the required False Alarm Rate of 2.6% to achieve HSS=0.50 for the imbalanced data set.

close to zero so that the number of false alarms does not exceed the number of hits. This is extremely challenging in SEP climatology; for example, only 37 events occurred in Solar Cycle 25 by the end of December 31, 2024. A model would need to achieve a similar number of false alarms across a 4 year period to obtain a high Threat Score and HSS. An even smaller number of false alarms, coupled with a high hit rate, is desired before SRAG could consider using a model for decision-making in operations. A new metric, called False Alarm Event Ratio, described in Section A.1, is proposed to capture the important aspect of desirable model performance.

In summary:

- The numerical values of many metrics cannot be compared across validation data sets with different levels of imbalance, e.g. SEPVAL and the SEP Scoreboards.
- Models forecasting in real time and evaluated for the true climatology of SEP events must generate much lower False Alarm Rates than those validated using balanced historical data sets to achieve comparable HSS and other skill scores.
- HSS, Threat Score, Bias, and False Alarm Ratio are sensitive to the relative numbers of false alarms compared to hits while TSS and False Alarm Rate are sensitive to the ratio of false alarms to clear periods. SRAG would like to attain a high Hit Rate while minimizing the number of false alarms, therefore the former group of metrics is most appropriate for SRAG's needs.

3.5 Forecast Deoverlapping

Many SEP models produce forecasts on a regular cadence. MAG4 and MagPy produce forecasts every hour as new magnetograms become available. SAWS-ASPECS forecast module produces forecasts every 3 hours depending on magnetogram availability. UMASEP monitors the *in situ* proton flux and issues a new forecast with each new GOES data point, every 5 minutes. For SEPVAL (but not on the SEP

Scoreboards), SPRINTS issues a new forecast for every minute while a flare is ongoing, updating the forecast as new X-ray measurements become available. In all cases, the forecast prediction window is longer than the cadence at which forecasts are issued. This results in multiple overlapping predictions for any given point in time. In SPHINX, each forecast is evaluated and counted individually. In some cases, this can produce metrics that are hard to interpret, particularly for All Clear.

As an example, UMASEP might issue tens of forecasts that overlap with an SEP event threshold crossing or dozens of false alarms over just a few hours. The contingency table for every individual UMASEP-10 forecast calculated for the lifetime of the SEP Scoreboards is shown in Table 3.4 (left). There were 37 SEP events since the start of the SEP Scoreboards, but the contingency table shows 1138 hits and a nearly equal number of 1131 misses. What does this mean in terms of individual SEP events? The contingency table also show 8711 false alarms and 797,871 correct negatives. The false alarms make up only 1.1% of all forecasts during quiet periods, but on the other hand, 8711 false alarms sound very disruptive in an operational environment. An additional complication, mentioned in Section 3.2.6, is that forecasts during SEP events are not evaluated by SPHINX. This has the effect of removing forecasts for the varying durations of SEP events and essentially applies a weighting that de-emphasizes SEP events and emphasizes quiet periods, the opposite of what is desired.

A method to "deoverlap" forecasts was developed to calculate a single validation result for all forecasts within a given period of time. For this report, deoverlapping was applied to calculate All Clear metrics. The deoverlapped All Clear outcomes were calculated as follows:

- Forecasts associated with an SEP event: Any Hit = Hit
- Forecasts associated with an SEP event: All Misses = Miss
- Forecasts associated with a non-event period: Any False Alarm = False Alarm
- Forecasts associated with a non-event period: All Correct Negatives = Correct Negative

The scheme developed for deoverlapping All Clear emphasizes both hits and false alarms, in line with SRAG's needs. For SEPVAL, all forecasts associated with each challenge period were combined to create a single result for each period. For the SEP Scoreboards, a single result was identified for each SEP event and each 24 hour quiet period — essentially the model forecasts were translated to daily forecasts.

Following this approach, the corresponding deoverlapped contingency table for UMASEP-10 for the SEP Scoreboards is shown in Figure 3.4 (right). Now we are able to interpret that UMASEP-10 has provided 1,361 days of forecasts to the SEP Scoreboards. It forecasted at least one hit for 24 SEP events and completely missed 11. On quiet days, 1,285 were correctly predicted as clear while 41 days had at least one false alarm. We now see that the 8,711 forecasts were not generated at random times, but were constrained to 41 days out of 1,361, or 3% of the days that it has been forecasting on the SEP Scoreboards. These false alarm days are roughly equal

to the 35 days with SEP events, much smaller than the ratio of nearly eight times more false alarms as reported in the raw scores.

Individual Forecasts

	Observed		
	Yes	No	Sum
Pred. Yes	1138	8711	9849
Pred. No	1131	797,871	799,002
Sum	2269	806,582	808,851

Deoverlapped

	Observed		
	Yes	No	Sum
Pred. Yes	24	41	65
Pred. No	11	1285	1296
Sum	35	1326	1361

Table 3.4: Contingency tables for UMASEP-10 for the total number of individual forecasts submitted to the SEP Scoreboards (left) and the deoverlapped results for 24 hour periods (right).

For SEPVAL, deoverlapping was applied to MAG4, MagPy, UMASEP, and SPRINTS. For the SEP Scoreboards, deoverlapped All Clear results are reported for GSU, MAG4, MagPy, and UMASEP.

Note that deoverlapping will tend to improve a model's All Clear scores if false alarms display some kind of pattern. This is the case for UMASEP, which tends to produce many false alarms in quick succession for a short period of time, discussed further in Section 6.11. If a model produces false alarms in a more random manner, deoverlapping tends to have less effect on the scores.

3.6 GOES Proton Observational Data

The validation results presented here were calculated with respect to GOES integral proton fluxes provided by National Oceanic and Atmospheric Administration (NOAA). GOES-15 and previous satellite measurements were downloaded from the NOAA NESDIS archives[8] for dates prior to March 2020. From March 2020 to the present, real-time GOES integral fluxes served by NOAA SWPC and archived at CCMC in the iSWA database[9] were used. These fluxes were downloaded and processed using FetchSEP[10], described in Section 2.1, to extract threshold crossing times and additional proton information for quiet periods and individual SEP events.

The low-energy GOES proton channels are known to suffer from contamination during SEP events when particles with energies above 50 - 60 MeV are present (Bruno, 2017; Posner, 2007). These particles can penetrate the GOES/EPEAD (or older GOES/EPS) detector shielding and create a spurious signal in the lower energy differential channels and the >10 MeV integral channel. This contamination has the effect of producing a spurious rise in the >10 MeV proton intensities at the beginning of an SEP event, in some cases resulting in an earlier-than-realistic threshold crossing time. The contamination may also artificially inflate the overall

[8] https://www.ncei.noaa.gov/data/goes-space-environment-monitor/access/avg/
[9] https://iswa.gsfc.nasa.gov/IswaSystemWebApp/hapi/
[10] https://github.com/ktindiana/fetchsep

>10 MeV proton intensities near the peak of the event if very energetic particles are present. These effects have not yet been quantified, but studies are underway within SRAG, working with various collaborators, to better understand the impact of this contamination and to produce a "cleaner" proton dataset. When such a dataset becomes available, the validation results can be recalculated using the SPHINX Validation Framework.

This analysis focuses on evaluating forecasts for All Clear threshold crossed/not-crossed outcomes, the probability of occurrence, and peak flux. Despite contamination effects, the observed binary outcome of whether >10 MeV flux exceeds 10 pfu should be robust. For the onset peak or maximum flux comparisons, the validation analysis presented here considers forecasts that fall within an order of magnitude of the observed value to be "good", leaving leeway for contamination in the measurements. The AWT is calculated with respect to the >10 MeV threshold crossing time, which may be incorrect due to contamination. However, this threshold crossing time with respect to the GOES data is used by SRAG in their support of Mission Control and reflects the timing relevant for making operational decisions. Therefore, it is meaningful to evaluate how models perform with respect to this value.

In summary:

- Proton fluxes measured by NOAA's GOES series of spacecraft suffer from contamination in the lower energy channels when protons above 50 - 60 MeV are present as they penetrate the detector shielding. This contamination creates unphysical onsets at the start of SEP events in the >10 MeV channel and may affect peak flux values.

- GOES spacecraft are the operational source of proton fluxes used in SRAG operations to support NASA Mission Control, therefore SRAG is interested in SEP model performance with respect to GOES measurements.

- SPHINX was built to be flexible so that validation may be performed with respect to any satellite measurements. If and when a higher quality dataset becomes available for the GOES proton fluxes, the validation results can be readily recalculated with SPHINX.

4 Validation Datasets

It should be noted that, throughout this report, we will typically use the generic term Solar Energetic Particles (SEP), however the terms Solar Particle Event (SPE) and Energetic Solar Particle Event (ESPE) carry specific operational meaning for SRAG and are used at times.

- SPE: >10 MeV protons exceed 10 pfu, relevant to astronauts during EVA
- ESPE: >100 MeV protons exceed 1 pfu, relevant to astronauts in a spacecraft

4.1 SEPVAL Dataset

The SEPVAL challenge contains 33 SEP events and a nearly equal number of 30 non-event periods to enable a validation of each model that tests for hits, misses, false alarms and correct negatives. The so-called "SEPVAL 2023 Challenge" lists are available on Zenodo [11] and listed in Table 4.1 (non-event periods) and Table 4.2 (SEP events). Note that there were two strong flares and fast CMEs in quick succession for the large SEP event on 2012-03-07, so it is listed in Table 4.2 twice. Models were given the opportunity to produce forecasts using both eruptions as inputs. Also note that a start time is not listed for the event on 2017-09-06. The >10 MeV proton flux was already elevated from the SEP on 2017-09-05 and no significant new onset was visible in that channel on 2017-09-06. This event was not included in the >10 MeV All Clear evaluation but was included in peak flux comparisons.

Most of the event periods in the SEPVAL challenge are associated with strong M and X class flares and fast CMEs, as described in Section 3.1 and shown in Figures 3.2–3.5. For each CME in the challenge, M2M used all available data to check the quality of the fits reported in the DONKI catalog and update the fits as needed to ensure that all CME parameters were the highest quality possible. The updated CME parameters are saved in the "SEPVAL_CME_CATALOG" in DONKI[12].

The participating SEP models covered a wide range of model approaches and forecasted quantities (see Table 4.3) as well as observational inputs and number of forecasts provided (see Table 4.4). A main goal of the SEPVAL challenge was to evaluate models as they would perform without human intervention. Participants were asked to produce forecasts in a "real-time-like" mode without changing parameters, tuning, or recalibrating their models from event to event. Each developer was asked to indicate which event periods in the challenge list were used in training their model. SRAG operational energy channels and thresholds for SEP and ESPE were solicited (but not required).

Forecasts were received for 20 different models from a wide variety of United States and international institutions, including universities and operational entities. The participating models are Air Force Dynamic Energetic Particle Tool (ADEPT), COronal Mass Ejections and Solar Energetic Particles (COMESEP), cRT+AE10, Solar Energetic Particle MODel (SEPMOD), Lavasa model, MagPy, Multivariate Ensemble of Models for Probabilistic SEP prediction (MEMPSEP), Proton Prediction System (PPS), Multiple Field Line Advection Model for Particle Acceleration (M-FLAMPA), SEP Advanced Warning System (SAWS)-Advanced Solar Particle Events Casting System (ASPECS), SEPSAT, SEPSTER, SEPSTER2D, Solar Particle Radiation Environment Analysis and Forecasting – Acceleration and Scattering Transport (SPREAdFAST), Space Radiation Intelligence System (SPRINTS), SPE Threat Assessment Tool (STAT), University of Málaga Solar Energetic Particle (UMASEP), UNSPELL, and improved Particle Acceleration and Transport in the Heliosphere (iPATH).

[11] https://doi.org/10.5281/zenodo.15020584
[12] https://kauai.ccmc.gsfc.nasa.gov/DONKI/search/

The models span a wide range of purposes and maturity levels. Some models, like STAT and M-FLAMPA, were created to understand the details of the physics behind solar energetic particle production and transport. Other models, like COMESEP, PPS, and UMASEP, have been forecasting in operational settings for years. Some of models, like cRT+AE10, MEMPSEP, and UNSPELL, were very recently developed, exploring new machine learning (ML) approaches for forecasting SEPs.

For the first time, the SEPVAL challenge evaluated a wide variety of SEP prediction models for the same set of benchmark events, testing their ability to discriminate between conditions that lead to SEP events and those that do not. There are more SEP models in the operations and research communities—Whitman et al. (2023) summarized 36 SEP models and new models are in development—however, the breadth of approaches, purposes, maturity levels, forecasted quantities, and developers make the participating models a representative sample. The SEPVAL challenge is the first assessment of the current state-of-the-art performance of SEP forecasting.

In summary:

- The SEPVAL community validation challenge received forecasts from 20 different SEP prediction models in the research community.
- The SEPVAL results represent the first assessment of the state-of-the-art performance of SEP model forecasting capabilities.

SEPVAL Non-Event Period	Flare Start	Flare Class	CME Start	CME Speed [km/s]	CME Width [deg]
2011-05-09	2011-05-09 20:42	C5.4	2011-05-09 20:57	776.0	58.0
2012-03-04	2012-03-04 10:29	M2.0	2012-03-04 11:00	1540.0	60.0
2012-03-05	2012-03-05 03:30	X1.1	2012-03-05 04:00	1363.0	51.0
2012-06-13	2012-06-13 11:29	M1.2	2012-06-13 14:00	643.0	39.0
2012-06-29	2012-06-29 09:13	M2.2	2012-06-29 09:36	1344.0	23.0
2013-06-07	2013-06-07 22:32	M5.9	2013-06-07 23:12	704.0	39.0
2013-06-28	2013-06-28 01:36	C4.4	2013-06-28 02:00	1200.0	41.0
2014-08-01	2014-08-01 18:00	M1.5	2014-08-01 18:36	709.0	36.0
2014-10-24	2014-10-24 07:37	M4.0	2014-10-24 08:00	728.0	19.0
2014-11-06	2014-11-06 03:32	M5.4	2014-11-06 04:00	662.0	38.0
2014-11-07	2014-11-07 16:53	X1.6	2014-11-07 18:08	802.0	52.0
2014-12-17	2014-12-17 04:25	M8.7	2014-12-17 05:00	680.0	26.0
2014-12-18	2014-12-18 21:41	M6.9	2014-12-19 01:04	1405.0	44.0
2015-03-09	2015-03-09 23:29	M5.8	2015-03-10 00:00	1711.0	22.0
2016-07-23	2016-07-23 05:00	M7.6	2016-07-23 05:24	740.0	40.0
2021-11-01	2021-11-01 00:57	M1.5	2021-11-01 02:00	657.0	47.0
2021-11-02	2021-11-02 02:03	M1.6	2021-11-02 02:48	1151.0	51.0
2022-01-18	2022-01-18 17:01	M1.5	2022-01-18 17:48	962.0	47.0
2022-04-17	2022-04-17 03:17	X1.1	2022-04-17 03:48	841.0	47.0
2022-04-20	2022-04-20 03:41	X2.2	2022-04-20 04:12	868.0	30.0
2022-04-29	2022-04-29 07:15	M1.2	2022-04-29 07:36	1397.0	48.0
2022-05-25	2022-05-25 18:12	M1.3	2022-05-25 18:36	929.0	34.0
2022-08-17	2022-08-17 13:26	M2.0	2022-08-17 14:36	767.0	49.0
2022-08-18	2022-08-18 10:37	M1.5	2022-08-18 11:00	1076.0	35.0
2022-08-19	2022-08-19 04:14	M1.6	2022-08-19 04:49	878.0	47.0
2022-08-29	2022-08-29 16:15	M3.8	2022-08-29 17:00	949.0	30.0
2022-08-30	2022-08-30 18:05	M2.1	2022-08-30 18:12	1378.0	32.0
2022-12-01	2022-12-01 07:04	M1.0	2022-12-01 07:48	1306.0	30.0
2023-03-04	2023-03-04 15:19	M5.2	2023-03-04 15:36	894.0	45.0
2023-03-06	2023-03-06 02:08	M5.8	2023-03-06 03:12	970.0	35.0

Table 4.1: SEPVAL Challenge non-events. The CME Start indicates the LASCO First Look Time. The CME speed and half-width are 3D measurements provided by M2M and recorded in DONKI.

SEPVAL SEP Event Period	Flare Start	Flare Class	CME Start	CME Speed [km/s]	CME Width [deg]	>10 MeV SEP Start	>100 MeV SEP Start
2011-03-08	2011-03-07 19:43	M3.7	2011-03-07 20:00	1980	45	2011-03-08 01:05	
2011-06-07	2011-06-07 06:16	M2.5	2011-06-07 06:49	1400	46	2011-06-07 08:20	2011-06-07 07:20
2011-08-04	2011-08-04 03:41	M9.3	2011-08-04 04:12	1950	60	2011-08-04 06:35	2011-08-04 05:10
2011-08-09	2011-08-09 07:48	X6.9	2011-08-09 08:12	1175	20	2011-08-09 08:45	2011-08-09 08:25
2012-01-23	2012-01-23 03:38	M8.7	2012-01-23 04:00	2211	62	2012-01-23 05:30	2012-01-23 04:45
2012-01-27	2012-01-27 17:37	X1.8	2012-01-27 18:27	2200	55	2012-01-27 19:05	2012-01-27 19:00
2012-03-07	2012-03-07 00:02	X5.4	2012-03-07 00:24	2809	54	2012-03-07 05:10	2012-03-07 04:05
2012-03-07	2012-03-07 01:05	X1.3	2012-03-07 01:30	2040	50	2012-03-07 05:10	2012-03-07 04:05
2012-03-13	2012-03-13 16:21	M7.9	2012-03-13 17:36	2250	60	2012-03-13 18:10	2012-03-13 18:10
2012-05-17	2012-05-17 01:25	M5.1	2012-05-17 01:48	1263	54	2012-05-17 02:10	2012-05-17 02:00
2012-07-07	2012-07-06 23:01	X1.1	2012-07-06 23:24	1200	40	2012-07-07 04:00	
2012-07-12	2012-07-12 15:37	X1.4	2012-07-12 16:48	1400	70	2012-07-12 18:35	
2012-07-23	Beyond West Limb		2012-07-23 02:36	2395	54	2012-07-23 15:45	
2012-09-28	2012-09-27 23:36	C3.7	2012-09-28 00:12	1252	47	2012-09-28 03:00	
2013-04-11	2013-04-11 06:55	M6.5	2013-04-11 07:24	743	48	2013-04-11 10:55	2013-04-11 09:40
2013-05-22	2013-05-22 12:30	M5.0	2013-05-22 13:25	1756	64	2013-05-22 14:35	2013-05-22 14:35
2013-09-30	2013-09-29 21:43	C1.2	2013-09-29 22:12	1100	70	2013-09-30 05:05	
2014-01-06	Beyond West Limb		2014-01-06 08:00	1138	51	2014-01-06 09:15	2014-01-06 08:30
2014-01-07	2014-01-07 18:04	X1.2	2014-01-07 18:24	2048	50	2014-01-07 19:20	2014-01-07 20:30
2014-02-25	2014-02-25 00:41	X4.9	2014-02-25 01:25	1670	66	2014-02-25 14:10	
2014-04-18	2014-04-18 12:31	M7.3	2014-04-18 13:25	1244	47	2014-04-18 15:25	
2014-09-11	2014-09-10 17:21	X1.6	2014-09-10 18:00	1400	45	2014-09-11 02:55	
2015-10-29	Beyond West Limb		2015-10-29 02:36	535	20	2015-10-29 05:50	2015-10-29 04:35
2017-07-14	2017-07-14 01:07	M2.4	2017-07-14 01:25	750	49	2017-07-14 09:00	
2017-09-05	2017-09-04 20:28	M5.5	2017-09-04 20:36	1323	54	2017-09-05 00:40	
2017-09-06	2017-09-06 11:53	X9.3	2017-09-06 12:24	1636	45		
2017-09-10	2017-09-10 15:35	X8.2	2017-09-10 16:00	2314	58	2017-09-10 16:45	2017-09-10 16:25
2021-05-29	2021-05-28 22:19	C9.4	2021-05-28 23:12	949	44	2021-05-29 03:00	
2021-10-28	2021-10-28 15:17	X1.0	2021-10-28 15:48	1109	49	2021-10-28 17:40	2021-10-28 16:35
2022-01-20	2022-01-20 05:41	M5.5	2022-01-20 06:12	1426	44	2022-01-20 08:00	2022-01-20 07:45
2022-03-28	2022-03-28 10:58	M4.0	2022-03-28 12:00	662	45	2022-03-28 13:25	2022-03-28 12:45
2022-04-02	2022-04-02 12:56	M3.9	2022-04-02 13:36	1370	45	2022-04-02 14:40	
2022-08-27	2022-08-27 01:52	M4.8	2022-08-27 02:24	1372	40	2022-08-27 11:55	
2023-02-25	2023-02-25 18:40	M6.3	2023-02-25 19:24	920	58	2023-02-25 21:10	

Table 4.2: SEPVAL Challenge SEP events. The CME Start indicates the LASCO First Look Time. The CME speed and half-width are 3D measurements provided by M2M and recorded in DONKI.

SEPVAL Model	Method	Energy Channels (MeV)	Forecast Quantities
ADEPT 1hr, 6hr	Empirical	>10	Time Profile
COMESEP flare	Empirical	>10	Probability, Peak
COMESEP flare+CME	Empirical	>10	Probability, Peak
cRT+AE10	ML	>10	Probability
ENLIL+SEPMOD	Physics-based	>10, >30, >50, >100	Time Profile
Lavasa	ML	>10	All Clear
MAG4	Empirical	>10	Probability
MagPy	Empirical	>10	Probability
MEMPSEP Mean, Median	ML	>10	Probability
MFLAMPA	Physics-based	>10, >30, >50, >100	Time Profile
PPS (SFS Update)	Empirical	>10, >100	Peak Flux
SAWS-ASPECS	Empirical, Physics-based	>10, >100	Probability, Time Profile
SAWS-ASPECS electrons	Empirical, Physics-based	>10, >100	Probability, Time Profile
SEPSAT	Physics-based	>10, >100	Time Profile
SEPSTER	Empirical	>10, >30, >50, >100	Peak Flux
SEPSTER2D	Empirical	>10, >30, >50, >100	Peak, Fluence
SPREAdFAST	Physics-based	>10, >30, >50, >100	Time Profile
SPRINTS 0-24 hour	ML	>10, >30, >50, >100	Probability, Peak
STAT	Physics-based	>10, >30, >50, >100	Time Profile
UMASEP-10	ML, Empirical	>10	Peak, Start
UMASEP-100	ML, Empirical	>100	Peak, Start
UNSPELL	ML	>5	Probability
ZEUS+iPATH	Physics-based	>10, >30, >50, >100	Time Profile

Table 4.3: Model approach, predicted energy channels, and forecasted quantities for models participating in the SEPVAL challenge (in alphabetical order).

SEPVAL Model	Observational Inputs	Cadence	N Forecasts
ADEPT 1hr, 6hr	protons	1/SEP	25
COMESEP flare	X-ray, EUV	1/flare	60
COMESEP flare+CME	X-ray, EUV, CME	1/CME	63
cRT+AE10		1/eruption	63
ENLIL+SEPMOD	Magnetograms, CME	1/CME	63
Lavasa	X-ray, CME	1/flare	58
MAG4_LOS_r	Magnetograms, SWPC Solar Region Summary (SRS)	1 hour	509
MAG4_SHARP_HMI	Magnetograms, SWPC SRS	1 hour	1462
MagPy	Magnetograms, SWPC SRS	1 hour	2182
MEMPSEP Mean, Median	X-ray, magnetograms, solar wind, suprathermal, protons, electrons	1/flare	60, 60
MFLAMPA	Magnetograms, EUV, CME	1/CME	9
PPS (SFS Update)	X-ray, ground-based radio	1/flare	61
SEPSAT	Magnetograms, EUV, CME, solar wind	1/CME	64
SEPSTER	Solar wind, CME	1/CME	64
SEPSTER2D	Solar wind, CME	1/CME	60
SPREAdFAST	Magnetograms, EUV, CME, suprathermal, protons	1/CME	8
SPRINTS 0-24 hour	X-ray, magnetograms, EUV	1 minute during flare	15263
STAT	Magnetograms, EUV, white light, CME, suprathermal	1/CME	6
UMASEP-10	X-ray, protons, SWPC SRS	5 minutes	27572
UMASEP-100	X-ray, protons, SWPC SRS	5 minutes	32240
UNSPELL	X-ray	1/flare	61
ZEUS+iPATH	Magnetograms, EUV, CME, solar wind, suprathermal	1/CME	60
SAWS-ASPECS	X-ray, Magnetograms, EUV, CME, protons	1/flare, 1/CME	57 - 63
SAWS-ASPECS electrons	X-ray, Magnetograms, EUV, CME, protons	1/flare, 1/CME	57 - 63

Table 4.4: Observational inputs, model cadence, and the number of forecasts provided by each model for the SEPVAL challenge.

4.2 SEP Scoreboards Dataset

The first models onboarded into the SEP Scoreboards began continuously producing forecasts in real time starting in March 2020; see Figure 3.1 for a timeline. Additional models were added as their technological maturity increased through their collaborations with ISEP. The Scoreboard models are SWPC, Magnetogram Forecast (MAG4), MagPy, Georgia State University (GSU), SAWS-ASPECS, SPRINTS, SEPSTER, SEPSTER2D, iPATH, SEPMOD, UMASEP, High Energy Solar Particle Events foRecastIng and Analysis (HESPERIA)-Relativistic Electron Alert System for Exploration (REleASE).

Models and forecasted quantities currently available on the SEP Scoreboards are listed in Table 4.6. Model onboarding dates and the number of forecasts produced during the reporting period for this Technical Report are listed in Table 4.7. The wide variety of observational inputs, human-in-the-loop requirements, and filters applied in the forecasting schemes are listed in Table 4.8.

The Scoreboard models use inputs available in real time to produce their forecasts. These inputs include satellite particle and X-ray fluxes, white light and EUV imagery, catalogs provided by NOAA SWPC and other sources, and inputs provided by M2M. M2M provides critical human-in-the-loop activities required to run the SEP Scoreboards in real time, including measuring CME parameters and entering them into the DONKI[13] catalog where models automatically access them and issue new forecasts.

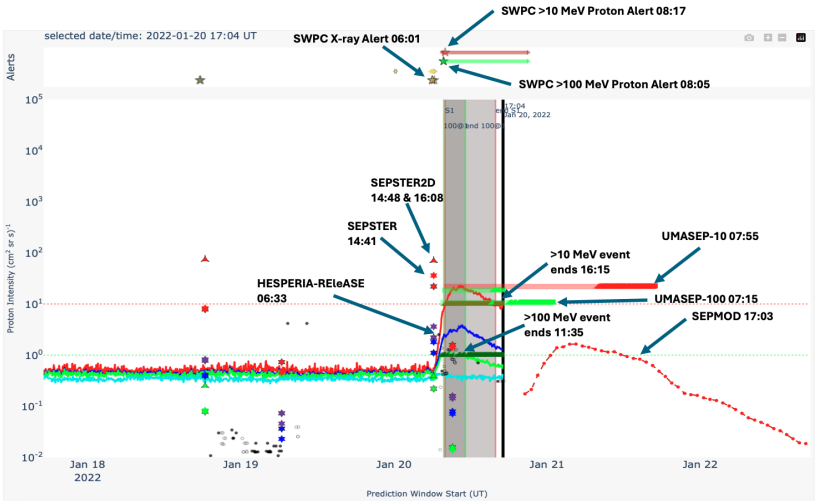
It is important to keep in mind that validation metrics for models on the SEP Scoreboards include:

- Model approach (internal theory and empirical, statistical relationships)
- Availability and quality of real time measurements (satellite and ground-based sources)
- Availability of human-in-the-loop analyses
- Availability of information served in real time in various catalogs (SWPC reports, Solarsoft, CACTus, etc)

The SEP Scoreboards are real-time systems used for applied research and environmental awareness. They run continuously in real time, but are not operational. They are being used by SRAG, which is an operational entity, for environmental awareness and to identify promising models, as well as for envisioning how SEP forecasting could be incorporated into operations or used as a support tool for astronauts. Models on the SEP Scoreboards may use measurements from scientific observatories that experience regular data gaps or that do not have a backup data source in the event of an unforeseen data outage (e.g. Solar Dynamics Observatory (SDO) data was unavailable for multiple months after a broken pipe flooded the Joint Science Operations Center (JSOC) computer center[14]). Models that require CME measurements from DONKI are subject to the availability of the Moon

[13] https://ccmc.gsfc.nasa.gov/tools/DONKI/

[14] Broken water pipe knocks out data processing for NASA sun-studying spacecraft, *Space.com*.



Scoreboard	UT	ET
X-ray alert	06:01	01:01
REleASE	06:31	01:31
UMASEP-100	07:15	02:15
UMASEP-10	07:55	02:55
>100 MeV start	08:05	03:05
>10 MeV start	08:15	03:15
>100 MeV end	11:35	06:35
SEPSTER	14:41	09:41
SEPSTER2D	14:47	09:47
SEPSTER2D rev	16:08	11:08
>10 MeV end	16:15	11:15
SEPMOD	17:03	12:03

Figure 4.1: Timing of SWPC alerts and model forecasts make the the Intensity SEP Scoreboard on January 20, 2022.

to Mars Space Weather Analysis Office which do not currently operate on a 24/7 schedule, *but will do so during Artemis missions.*

Figure 4.1 lists the timing of SWPC alerts and forecasts issued to the Intensity SEP Scoreboard as the January 20, 2022 SEP event unfolded, resulting in both >10 MeV and >100 MeV threshold crossings. The SEP event was produced by an M5.5 flare followed by a 1426 km/s CME. SWPC sent out an X-ray alert after the flare reached peak at 06:01 UT. Note that 06:01 UT is 1:01 am Eastern Time, the middle of the night for the M2M office. The SEP Scoreboard remained quiet until REleASE produced a forecast above threshold at 06:31 UT, followed by UMASEP-100 45 minutes later at 07:15 UT. UMASEP-10 issued a prediction 40 minutes later at 07:55 UT. The >100 MeV proton flux exceeded the 1 pfu operational threshold at 08:05 UT and the >10 MeV proton flux exceeded the 10 pfu operational threshold 10 minutes later at 08:15 UT (3:15 am ET). The remaining models on the Intensity SEP Scoreboard require CME measurements to trigger their forecasts. The SEP Scoreboard remained quiet and the >100 MeV ESPE ended at 11:35 UT (5:35 am ET) without any additional forecasts. Finally, at 14:41 UT (9:41 am ET), M2M analysts arrived at work and had the chance to review the space weather conditions overnight and fit the CME responsible for the ongoing SEP event. SEPSTER and SEPSTER2D forecasts are issued in quick succession at 14:41 UT and 14:47 UT, respectively. Two hours later, at 16:08 UT, M2M remeasured the CME parameters and submitted them to DONKI. Revised SEPSTER and SEPSTER2D forecasts were issued and an Wang-Sheeley-Arge Enlil (WSA-Enlil)+SEPMOD run was initiated. The >10 MeV SPE ends just 7 minutes later at 16:15 UT (11:15 am ET). Finally, WSA-Enlil+SEPMOD completed and issued a forecast to the SEP Scoreboard at 17:03 UT (12:03pm ET).

The above example demonstrates that the validation metrics and, particularly the AWT, reflect the availability and quality of real-time measurements, the underlying

model approach, and the SEP Scoreboard system as a whole.

The forecasts aggregated by the SEP Scoreboards are an extremely valuable dataset representing true real-time performance. The 11 different models participating in the Scoreboards, alongside SWPC forecasts, are the best representative sample available of our current real-time capabilities for SEP prediction.

In summary:

- The SEP Scoreboards went live in March of 2020, performing the critical role of collecting and visualizing forecasts produced in real time.
- Validation of the SEP Scoreboards represents true forecasting performance, including availability and quality of input data in real time, time for human-in-the-loop analyses, computational run-time, and inherent model approach.

Flare Start	Flare Class	CME Start	CME Speed [km/s]	CME Width [deg]	>10 MeV SEP Start	>100 MeV SEP Start
2021-05-28 22:19	C9.4	2021-05-28 23:12	824	46	2021-05-29 03:00	
2021-10-28 15:17	X1.0	2021-10-28 15:48	1109	49	2021-10-28 17:40	2021-10-28 16:35
2022-01-20 05:59	M5.5	2022-01-20 06:12	1426	44	2022-01-20 08:00	2022-01-20 07:20
2022-03-28 10:58	M4.0	2022-03-28 12:00	662	45	2022-03-28 13:25	2022-03-28 12:45
2022-03-30 17:21	X1.3	2022-03-30 18:00	808	40	2022-03-31 06:20	
2022-04-02 12:56	M3.9	2022-04-02 13:36	1370	45	2022-04-02 14:30	
2022-08-27 01:52	M4.8	2022-08-27 02:24	1372	40	2022-08-27 11:55	
2023-02-25 18:40	M6.3	2023-02-25 19:12	920	58	2023-02-25 21:10	
		2023-03-13 03:36	2127	50	2023-03-13 07:45	
2023-04-21 17:44	M1.7	2023-04-21 18:12	1204	43	2023-04-23 18:15	
2023-05-07 22:53	M1.6	2023-05-07 23:12	1494	44	2023-05-08 12:40	
2023-05-09 18:20	M4.2	2023-05-09 19:00	1651	45	2023-05-09 23:35	
		2023-07-16 04:42	1220	51	2023-07-16 05:45	
2023-07-17 23:20	M5.7	2023-07-17 23:36	1388	53	2023-07-18 01:15	
2023-07-28 15:39	M4.1	2023-07-28 16:00	2000	43	2023-07-29 00:20	
2023-08-05 06:16	M1.6	2023-08-05 07:00	1044	46	2023-08-05 11:15	
2023-08-05 21:45	X1.6	2023-08-05 22:24	1757	49	2023-08-05 23:45	
2023-08-07 20:30	X1.5	2023-08-07 20:48	1429	45	2023-08-08 01:15	
2023-09-01 02:54	M1.2	2023-09-01 03:24	1142	49	2023-09-01 04:30	
2023-12-15 07:23	M6.9	2023-12-15 07:48	1101	37	2023-12-15 23:45	
2023-12-31 21:36	X5.0	2023-12-31 22:00	2184	45	2024-01-03 20:05	
2024-01-29 03:54	M6.8	2024-01-29 04:38	1277	54	2024-01-29 06:15	
2024-02-09 12:53	X3.3	2024-02-09 13:23	1754	46	2024-02-09 15:30	
		2024-02-12 06:36	3180	42	2024-02-12 08:05	2024-02-12 10:15
		2024-02-14 04:09	2012	43	2024-02-14 09:10	
2024-03-15 00:59	C1.9	2024-03-15 02:10	1121	51	2024-03-15 20:50	
2024-03-23 00:58	X1.1	2024-03-23 01:25	1613	41	2024-03-23 04:00	
2024-05-10 06:27	X3.9	2024-05-10 07:12	1018	41	2024-05-10 13:35	
2024-05-11 01:10	X5.8	2024-05-11 01:36	1263	51	2024-05-11 02:10	2024-05-11 02:10
2024-05-13 08:48	M6.6	2024-05-13 09:12	1456	49	2024-05-13 14:00	
2024-06-08 01:23	M9.7	2024-06-08 01:53	1106	45	2024-06-08 02:55	2024-06-08 02:25
		2024-07-23 00:23	1299	45	2024-07-23 03:00	2024-07-23 02:05
2024-09-09 00:57	M1.0	2024-09-09 05:23	1452	45	2024-09-09 08:50	
2024-09-14 15:13	X4.5	2024-09-14 15:36	2040	45	2024-09-17 07:35	
2024-10-09 01:25	X1.8	2024-10-09 02:12	1509	45	2024-10-09 05:05	2024-10-09 03:50
2024-10-26 06:32	X1.8	2024-10-26 06:48	1523	53	2024-10-26 18:05	
		2024-11-21 18:12	1560	49	2024-11-21 19:25	2024-11-21 18:45

Table 4.5: SEP Scoreboard era SEP events. The CME Start indicates the LASCO First Look Time. The CME speed and half-width are 3D measurements provided by M2M and recorded in DONKI.

SEP Scoreboard Model	Method	Energy Channels (MeV)	Forecast Quantities
SWPC Day 1	Human Forecaster	>10, >100	Probability
SWPC Warning	Human Forecaster	>10, >100	All Clear
MAG4 LOS FEr	Empirical	>10	Probability
MAG4 LOS r	Empirical	>10	Probability
MAG4 SHARP FE	Empirical	>10	Probability
MAG4 SHARP HMI	Empirical	>10	Probability
MAG4 SHARP	Empirical	>10	Probability
MagPy	Empirical	>10	Probability
GSU	ML	>10	Probability
SAWS-ASPECS	Empirical, Physics-based	>10, >30, >100, >300	Probability, Time Profile
SPRINTS	ML	>10, >30, >50, >100	Probability, Peak
SEPSTER	Empirical	>10, >30, >50, >100	Peak Flux
SEPSTER2D	Empirical	>10, >30, >50, >100	Peak, Fluence
ZEUS+iPATH	Physics-based	>10, >30, >50, >100	Time Profile
ENLIL+SEPMOD	Physics-based	>10, >30, >50, >100	Time Profile
UMASEP-10,-30,-50,-100	ML, Empirical	>10, >30, >50, >100	Max in prediction window, Start
UMASEP-500	ML, Empirical	>500	Peak, Start
HESPERIA REleASE	Empirical	15.8-39.8, 28.2-50.1	Flux in 30, 60, 90 minutes

Table 4.6: Model approach, predicted energy channels, and forecasted quantities for models forecasting in real time on the SEP Scoreboards (in order discussed in this report).

SEP Scoreboard Model	Date Onboarded	Duration (Years)	Cadence	Number of Forecasts	Number of SEP Events
SWPC Day 1			1 day		
SWPC Warning	2021/05	3.5	as needed	52	29
MAG4 LOS Fer	2021/02	3.8	1 hour	28,027	35
MAG4 LOS r	2021/01	3.9	1 hour	28,040	35
MAG4 SHARP FE	2021/04	3.6	1 hour	19,214	34
MAG4 SHARP HMI	2021/04	3.6	1 hour	19,813	34
MAG4 SHARP	2021/04	3.6	1 hour	19,531	34
MagPy SHARP HMI CEA	2023/06	1.4	1 hour	11,892	23
GSU	2023/04	1.6	1 hour	12,857	24
SPRINTS (All)	2022/08	2.3	1/flare	30,076	30
SAWS-ASPECS (All)	2022/10	2.2	1/flare, 1/CME, 3 hours	493,566	22
SEPSTER (Parker Spiral)	2020/04	4.6	1/CME	3,899	31
SEPSTER (WSA-ENLIL)	2020/04	4.6	1/CME	3,660	31
SEPSTER2D	2021/06	3.5	1/CME	1,103	31
ZEUS + iPATH CME	2023/06	1.6	1/CME	869	18
ENLIL + SEPMOD	2021/04	3.7	1/CME	2,045	29
UMASEP-10	2020/03	4.8	3 minutes	810,485	35
UMASEP-30	2021/01	4.0	3 minutes	814,363	35
UMASEP-50	2021/01	4.0	3 minutes	813,793	21
UMASEP-100	2020/03	4.8	3 minutes	841,771	8
HESPERIA REleASE ACE 60-min	2020/03	4.8	5 minutes	184,802	26
HESPERIA REleASE SOHO 60-min	2020/03	4.8	5 minutes	382,676	22

Table 4.7: Model forecast statistics for the SEP Scoreboards. The number of SEP events listed are for the $>$ 10 MeV channel unless the model's predicted energy range is explicitly different, e.g. UMASEP-100.

SEP Scoreboard Model	Observational Inputs	Human-in-the-loop	Filters
SWPC Day 1	All SpWx info	SWPC forecaster	N/A
SWPC Warning	protons	SWPC forecaster	N/A
MAG4 LOS Fer	MDI LOS magnetogram, SRS	No	No
MAG4 LOS r	MDI LOS magnetogram, SRS, SWPC Events	No	M & X flares
MAG4 SHARP FE	MDI vector magnetogram, SRS	No	None
MAG4 SHARP HMI	HMI vector magnetogram, SRS, SWPC Events	No	M & X flares
MAG4 SHARP	MDI vector magnetogram, SRS, SWPC Events	No	M & X flares
MagPy SHARP HMI CEA	HMI vector magnetogram, SRS, SWPC Events	No	None
GSU	HMI vector magnetogram	No	None
SPRINTS Post Eruptive 0-24 hrs	X-ray, Solarsoft	No	
SAWS-ASPECS flare	Solarsoft	No	
SAWS-ASPECS flare 50%, 90%	Solarsoft, protons	No	
SEPSTER (Parker Spiral)	DONKI CME, solar wind speed	M2M analyst	$\geq$200 km/s, $\geq$10 deg
SEPSTER (WSA-ENLIL)	DONKI CME, WSA-ENLIL	M2M analyst	$\geq$200 km/s, $\geq$10 deg
SEPSTER2D	DONKI CME, solar wind speed	M2M analyst	$\geq$600 km/s, $\geq$20 deg
ZEUS+iPATH CME	DONKI CME, ZEUS	M2M analyst	$\geq$450 km/s, $\geq$30 deg
ENLIL+SEPMOD	DONKI CME, WSA-ENLIL	M2M analyst	$\geq$450 km/s, $\geq$30 deg
UMASEP-10	protons, X-rays, SWPC Events	No	$\geq$C1 (WCP), $\geq$M2 (SOD), $\geq$5.9 pfu (PCP)
UMASEP-30	protons, X-rays	No	$\geq$C9 (WCP), $\geq$0.76 pfu (PCP)
UMASEP-50	protons, X-rays	No	$\geq$M2 (WCP), $\geq$0.56 pfu (PCP)
UMASEP-100	protons, X-rays	No	$\geq$M3.5 (WCP), $\geq$0.74 pfu (PCP)
UMASEP-500	protons, X-rays	No	$\geq$X2.5 (WCP)
HESPERIA REleASE	ACE, SOHO electrons	No	None

Table 4.8: Observational inputs, whether a human is in-the-loop to generate forecasts, and any filters applied to inputs for each SEP Scoreboard model. Protons are from GOES. Models that use WSA-ENLIL and ZEUS need magnetograms as inputs into the solar wind simulations. SRS is the NOAA Solar Region Summary and SWPC Events is https://services.swpc.noaa.gov/json/edited_events.json.

5 Group Validation Results

Here we report the median and distribution of selected metrics for the group of models that participated in the SEPVAL challenge and the group of models active on the SEP Scoreboards. We further divide the models into two subgroups, "post-eruptive" and "pre-eruptive", which have distinctly different forecasting intentions and skill.

Post-eruptive models produce a forecast following a solar eruption and answer the question: "Will an SEP occur after THIS eruption?" These models use observational inputs that are produced by the eruption itself, like flares, CMEs, or enhancements in the *in situ* proton and electron flux. These models typically produce forecasts at a cadence of solar eruptions.

Pre-eruptive models answer the question: "Will an eruption and SEP occur in the next X hours based on current conditions?" These models use the current conditions in active regions in the photosphere and/or corona to predict whether an eruption and associated SEP event will occur. If an eruption has recently been observed on the Sun, then forecasts from these types of models are interpreted as a prediction for the NEXT possible eruption. These models use magnetograms and SWPC active region products as observational inputs and typically produce forecasts at a regular cadence, e.g., hourly or some other regular time period. Models in the pre-eruptive category do not benefit from information about a specific eruption as post-eruptive models do.

Pre- and post-eruptive models face different forecasting challenges so their validation scores have been separated in this report.

5.1 SEPVAL Group Results

Here we discuss the summary metrics for the 33 SEP events and 30 non-event periods chosen as benchmark events for the SEPVAL community challenge, making an approximately balanced data set with an equal number of positive (SEP) and negative (non-event) samples. These results can be viewed as an idealized evaluation of model performance using their default workflows. As discussed in 4.1, the inputs are of higher quality than real-time observations and do not suffer from data gaps, the sets of model forecasts are as complete as possible, and the extreme imbalance of true climatology is not taken into account here.

This section includes median metrics for All Clear, probability, onset peak, and maximum flux with separate scores for post-eruptive and pre-eruptive models as identified in Table 5.1. Definitions of the onset peak and maximum flux are described in Section 2.1 and shown in Figure 2.4. Box plots for each metric show the range and distribution of the individual model scores.

5.1.1 All Clear

Figures 5.1 and 5.2 show the All Clear outcomes for >10 MeV for all participating SEPVAL models that provided All Clear forecasts for SEP event and non-event periods, respectively. Figures 5.3 and 5.4 show the same for >100 MeV, 1 pfu. The

SEPVAL Model	Model Type	Included in Median Metrics
ADEPT AFRL 1-hr	post-eruptive	Peak
COMESEP flare only	post-eruptive	Probability, Peak
COMESEP flare+CME	post-eruptive	Probability, Peak
cRT+AE10	post-eruptive	Probability
Lavasa	post-eruptive	All Clear
MAG4_LOS_r	pre-eruptive	All Clear, Probability
MAG4_SHARP_HMI	pre-eruptive	All Clear, Probability
MagPy_SHARP_HMI_CEA (19%)	pre-eruptive	All Clear, Probability
MagPy_SHARP_HMI_CEA (7%)	pre-eruptive	All Clear, Probability
MEMPSEP Mean	post-eruptive	Probability
MEMPSEP Median	post-eruptive	Probability
SAWS-ASPECS CME (SOHO)	post-eruptive	All Clear, Probability
SAWS-ASPECS CME (SOHO) 50%, 90%	post-eruptive	Peak
SAWS-ASPECS CME (SOHO) electrons	post-eruptive	All Clear, Probability
SAWS-ASPECS CME (SOHO) electrons 50%, 90%	post-eruptive	All Clear, Peak
SAWS-ASPECS flare + CME (SOHO)	post-eruptive	All Clear, Probability
SAWS-ASPECS flare + CME (SOHO) 50%, 90%	post-eruptive	Peak
SAWS-ASPECS flare + CME (SOHO) electrons	post-eruptive	All Clear, Probability
SAWS-ASPECS flare + CME (SOHO) electrons 50%, 90%	post-eruptive	Peak
SAWS-ASPECS flare	post-eruptive	All Clear, Probability
SAWS-ASPECS flare 50%, 90%	post-eruptive	Peak
SAWS-ASPECS flare electrons	post-eruptive	All Clear, Probability
SAWS-ASPECS flare electrons 50%, 90%	post-eruptive	Peak
SEPSAT	post-eruptive	All Clear, Peak
SFS-Update	post-eruptive	All Clear, Probability, Peak
SPRINTS Post Eruptive 0-24 hrs	post-eruptive	All Clear, Probability
ENLIL+SEPMOD	post-eruptive	All Clear, Peak
SEPSTER (Parker Spiral)	post-eruptive	All Clear, Peak
SEPSTER2D	post-eruptive	All Clear, Peak
UMASEP-10, UMASEP-100	post-eruptive	All Clear, Peak
UNSPELL flare	post-eruptive	All Clear, Probability
ZEUS+iPATH CME	post-eruptive	All Clear, Peak

Table 5.1: Models included in SEPVAL median metrics for All Clear, probability, onset peak and max flux).

sums at the bottom and right sides of the figures are components of each model's contingency table. The SEPVAL >10 MeV dataset has an imbalance of 1:0.94 (32:30 events to non-events) and the >100 MeV dataset has an imbalance of 1:2.7 (17:46 events to non-events). As described in Section 3.5, MAG4, MagPy, SPRINTS, and UMASEP results have been deoverlapped to derive a single answer for each challenge period. A label of "No Data" indicates that a forecast was not provided. These No Data periods are left out of the metrics and do not count for or against the model.

In Figure 5.1, many models have a status of No Data for 2017-09-06 which is primarily due to the ambiguous nature of the event's onset since particle fluxes were already elevated at the time. Many models provided forecasts for this event, but differences in alignment of the timing parameters within SPHINX (e.g. prediction window with observation period) resulted in no association for some models, leaving this event out of the statistics. For three events, 2012-07-23, 2014-01-06, and 2015-10-29, the associated flare was beyond the West limb and was not observed at Earth. All of the models that required flare data as input could not issue forecasts, as is evidenced by the numerous outcomes of No Data in those rows. While these aren't counted as misses in the metrics, **it demonstrates how models that use flare inputs are limited by our current one-sided view of the Sun**.

SEP Events	MAG4 LOS_r	MAG4 SHARP HMI	MagPy SHARP HMI_CEA	MagPy SHARP HMI_CEA_7	SPRINTS Post Eruptive 0-24 hrs	UMASEP-10	Lavasa	SAWS-ASPECS CME (SOHO)	SAWS-ASPECS CME (SOHO) electrons	SAWS-ASPECS flare + CME (SOHO)	[illegible] ASPECS flare + CME (SOHO) electrons	SAWS-ASPECS flare	SAWS-ASPECS flare electrons	SEPMOD	SEPSAT	SEPSTER (Parker Spiral)	SEPSTER2D CME	SFS-Update	UNSPELL flare	ZEUS+iPATH CME	Total Hits	Total Misses	Total No Data
2011-03-08 01:05:00	Hit	Hit	Miss	Hit	Hit	Hit	Hit	Hit	Hit	Hit	Hit	Miss	Miss	Hit	Hit	Hit	Hit	Hit	Hit	Hit	17	3	0
2011-06-07 08:20:00	Miss	Miss	Miss	Miss	Miss	Hit	Hit	Hit	Hit	Hit	Hit	Miss	Miss	Hit	Hit	Hit	No Data	Hit	Hit	Hit	12	7	1
2011-08-04 06:35:00	Hit	Hit	Miss	Hit	Hit	Hit	Hit	Hit	Hit	Hit	Hit	Hit	Hit	Hit	Hit	Hit	Hit	Hit	Hit	Hit	19	1	0
2011-08-09 08:45:00	Miss	Miss	Miss	Hit	Hit	Hit	Hit	Hit	Hit	Hit	Hit	Hit	Hit	Miss	Hit	Miss	Hit	Hit	Hit	Miss	14	6	0
2012-01-23 05:30:00	Hit	Hit	Miss	Miss	Hit	Hit	Hit	Hit	Hit	Hit	Hit	Hit	Hit	Hit	Hit	Hit	Hit	Hit	Hit	Hit	18	2	0
2012-01-27 19:05:00	Miss	Miss	Miss	Hit	Hit	Hit	Hit	Hit	Hit	Hit	Hit	Hit	Hit	Hit	Hit	Hit	Hit	Hit	Hit	Hit	17	3	0
2012-03-07 05:10:00	Hit	Hit	Miss	Hit	Hit	Hit	Hit	Hit	Hit	Hit	Hit	Hit	Hit	Hit	Hit	Hit	Hit	Hit	Hit	Miss	18	2	0
2012-03-13 18:10:00	Hit	Hit	Miss	Hit	Hit	Hit	No Data	Hit	Hit	Hit	Hit	Hit	Hit	Hit	Hit	Hit	Hit	Hit	No Data	Hit	17	1	2
2012-05-17 02:10:00	Hit	Miss	Miss	Miss	Miss	Hit	No Data	Hit	Hit	Hit	Hit	Hit	Hit	Hit	Miss	Hit	Hit	Hit	No Data	Hit	13	5	2
2012-07-07 04:00:00	Hit	Hit	Hit	Hit	Miss	Hit	Hit	Hit	Hit	Hit	Hit	Hit	Hit	Miss	Miss	Hit	Hit	Hit	Miss	Hit	16	4	0
2012-07-12 18:35:00	Hit	Hit	Miss	Hit	Hit	Hit	Hit	Hit	Hit	Hit	Hit	Hit	Hit	Miss	Hit	Hit	Hit	Hit	Hit	Hit	18	2	0
2012-07-23 15:45:00	Miss	Miss	Miss	Miss	No Data	Hit	No Data	Hit	Hit	No Data	No Data	No Data	No Data	Miss	Hit	Hit	Hit	No Data	No Data	Hit	7	5	8
2012-09-28 03:00:00	Miss	Miss	Miss	Miss	Miss	Hit	Miss	Hit	Hit	Miss	Miss	Miss	Miss	Hit	Hit	Hit	Hit	Miss	Miss	Hit	8	12	0
2013-04-11 10:55:00	Hit	Miss	Miss	Miss	Hit	Hit	Miss	Hit	Miss	Hit	Miss	Hit	Miss	Miss	Hit	Miss	Miss	Hit	Hit	Miss	9	11	0
2013-05-22 14:35:00	Hit	Miss	Miss	Miss	Hit	Hit	No Data	Hit	Hit	Hit	Hit	Hit	Hit	Hit	Hit	Hit	Hit	Hit	No Data	Miss	14	4	2
2013-09-30 05:05:00	Miss	Miss	Miss	Hit	Miss	No Data	No Data	Hit	Hit	Miss	Miss	Miss	Miss	Hit	Hit	Miss	Hit	Miss	Miss	Hit	7	11	2
2014-01-06 09:15:00	Hit	Hit	Miss	Hit	No Data	Hit	No Data	Hit	Hit	No Data	No Data	No Data	No Data	Miss	Hit	Miss	Hit	No Data	No Data	Miss	8	4	8
2014-01-07 19:20:00	Hit	Hit	Miss	Hit	Hit	Hit	Hit	Hit	No Data	Hit	No Data	Hit	No Data	Miss	Hit	Hit	Hit	Hit	Hit	Hit	15	2	3
2014-02-25 14:10:00	Hit	Hit	Hit	Hit	Miss	Hit	Miss	Hit	Hit	Hit	Hit	Hit	Hit	Miss	Hit	Miss	Hit	Hit	Hit	Miss	15	5	0
2014-04-18 15:25:00	Hit	Hit	Miss	Miss	Hit	Hit	Hit	Hit	Hit	Hit	Hit	Hit	Hit	Hit	Hit	Hit	Hit	Hit	Hit	Miss	17	3	0
2014-09-11 02:55:00	Hit	Hit	Miss	Miss	Hit	Hit	Hit	Hit	Hit	Hit	Hit	Hit	Hit	Hit	Hit	Hit	Hit	Hit	Hit	Hit	18	2	0
2015-10-29 05:50:00	Miss	Miss	Miss	Hit	No Data	Hit	No Data	Miss	Miss	No Data	No Data	No Data	No Data	Miss	Hit	Miss	Miss	No Data	No Data	No Data	3	8	9
2017-07-14 09:00:00	Miss	Miss	Miss	Miss	Hit	Hit	Hit	Hit	Hit	Hit	Hit	Miss	Miss	Miss	Hit	Miss	No Data	Hit	Hit	Miss	10	9	1
2017-09-05 00:40:00	Hit	Hit	Miss	Hit	Hit	Hit	Miss	Hit	Hit	Hit	Hit	Hit	Hit	Miss	Hit	Hit	Hit	Hit	Miss	Hit	16	4	0
2017-09-06 12:00:00	Hit	Hit	Hit	Hit	Miss	No Data	No Data	No Data	No Data	No Data	No Data	No Data	No Data	No Data	No Data	No Data	No Data	No Data	No Data	No Data	4	1	15
2017-09-10 16:45:00	Miss	Miss	Miss	Hit	Hit	Hit	Hit	Hit	Hit	Hit	Hit	Hit	Hit	Hit	Hit	Hit	Hit	Hit	Hit	Hit	17	3	0
2021-05-29 03:00:00	Miss	Miss	Miss	Hit	Miss	Miss	Hit	Miss	Miss	Miss	Miss	Miss	Miss	Hit	Miss	Miss	Hit	Hit	Hit	Hit	7	13	0
2021-10-28 17:40:00	Hit	Miss	Miss	Miss	Hit	Hit	Hit	Hit	Hit	Hit	Hit	Hit	Hit	Hit	Miss	Miss	Hit	Hit	Hit	Miss	14	6	0
2022-01-20 08:00:00	Miss	Miss	Miss	Hit	Miss	Hit	Miss	Hit	Hit	Hit	Hit	Hit	Hit	Miss	Hit	Hit	Hit	Hit	Hit	Miss	13	7	0
2022-03-28 13:25:00	Hit	Miss	Miss	Hit	Miss	Hit	Miss	Miss	Miss	Hit	Hit	Miss	Miss	Miss	Miss	Miss	Miss	Hit	Miss	Miss	6	14	0
2022-04-02 14:40:00	Hit	Miss	Miss	Hit	Hit	Hit	No Data	Hit	Hit	Hit	Hit	Miss	Miss	Hit	Hit	Hit	Hit	Hit	No Data	Hit	14	4	2
2022-08-27 11:55:00	Hit	Hit	Hit	Hit	Hit	Miss	Hit	Hit	Hit	Hit	Hit	Hit	Hit	Hit	Hit	No Data	Hit	Hit	Hit	Hit	18	1	1
2023-02-25 21:10:00	Hit	Hit	Miss	Hit	Hit	Hit	Hit	Hit	Hit	Hit	Hit	Hit	Hit	Hit	Hit	Miss	Hit	Hit	Hit	Hit	18	2	0
Total Hits	22	16	4	22	20	29	18	29	27	26	24	21	19	19	27	20	27	27	20	20			
Total Misses	11	17	29	11	10	2	6	3	4	3	4	8	9	13	5	11	3	2	5	11			
Total No Data	0	0	0	0	3	2	9	1	2	4	5	4	5	1	1	2	3	4	8	2			

Figure 5.1: All Clear outcomes for SEPVAL challenge >10 MeV, 10 pfu SEP events showing hit, miss, or no data for each model.

Non-Event Start	Non-Event End	MAG4 LOS_r	MAG4 SHARP HMI	MagPy SHARP HMI_CEA	MagPy SHARP HMI_CEA_7	SPRINTS Post Eruptive 0-24 hrs	UMASEP-10	Lavasa	SAWS-ASPECS CME (SOHO)	SAWS-ASPECS CME (SOHO) electrons	SAWS-ASPECS flare + CME (SOHO)	SAWS-ASPECS flare + CME (SOHO) electrons	SAWS-ASPECS flare	SAWS-ASPECS flare electrons	SEPMOD	SEPSAT	SEPSTER (Parker Spiral)	SEPSTER2D CME	SFS-Update	UNSPELL flare	ZEUS+iPATH CME	Total Correct Negatives	Total False Alarms	Total No Data
2011-05-08 21:42:00	2011-05-10 10:42:00	FA	CN	CN	CN	CN	CN	CN	FA	CN	CN	CN	CN	CN	CN	CN	CN	CN	CN	CN	CN	18	2	0
2012-03-03 11:29:00	2012-03-04 19:00:00	No Data	No Data	CN	FA	CN	CN	CN	FA	CN	FA	CN	CN	CN	CN	FA	CN	FA	FA	CN	CN	12	6	2
2012-03-04 19:00:00	2012-03-05 17:30:00	No Data	No Data	CN	FA	FA	CN	FA	FA	CN	FA	CN	FA	CN	CN	FA	CN	FA	FA	FA	CN	8	10	2
2012-06-12 12:29:00	2012-06-14 01:29:00	No Data	FA	CN	CN	FA	CN	CN	CN	CN	CN	CN	CN	CN	CN	CN	CN	CN	FA	CN	CN	16	3	1
2012-06-28 10:13:00	2012-06-29 23:13:00	FA	CN	CN	CN	CN	CN	CN	CN	CN	CN	CN	CN	CN	CN	FA	CN	FA	CN	CN	CN	17	3	0
2013-06-06 23:32:00	2013-06-08 12:32:00	CN	CN	CN	CN	CN	CN	CN	CN	CN	FA	CN	FA	CN	CN	CN	CN	CN	FA	FA	CN	16	4	0
2013-06-27 02:36:00	2013-06-28 15:36:00	CN	CN	CN	CN	CN	CN	CN	FA	CN	CN	CN	CN	CN	CN	CN	FA	FA	CN	CN	FA	16	4	0
2014-07-31 19:00:00	2014-08-02 08:00:00	No Data	No Data	CN	FA	CN	CN	CN	CN	CN	CN	CN	CN	CN	CN	CN	CN	CN	CN	CN	No Data	16	1	3
2014-10-23 08:37:00	2014-10-24 21:37:00	FA	FA	CN	FA	CN	FA	CN	CN	CN	CN	CN	CN	CN	CN	CN	CN	CN	FA	CN	CN	15	5	0
2014-11-05 04:32:00	2014-11-06 17:32:00	No Data	No Data	CN	FA	CN	CN	CN	CN	CN	FA	CN	FA	CN	CN	FA	CN	CN	FA	CN	CN	13	5	2
2014-11-06 17:53:00	2014-11-08 06:53:00	No Data	No Data	CN	FA	CN	CN	CN	CN	CN	FA	CN	FA	CN	CN	CN	CN	CN	FA	FA	CN	13	5	2
2014-12-16 05:25:00	2014-12-17 18:25:00	No Data	No Data	CN	CN	FA	CN	CN	CN	CN	FA	CN	FA	CN	CN	CN	CN	CN	FA	FA	CN	13	5	2
2014-12-17 22:41:00	2014-12-19 11:41:00	FA	FA	CN	FA	FA	CN	FA	FA	CN	FA	CN	FA	CN	CN	FA	CN	FA	FA	FA	CN	8	12	0
2015-03-09 00:00:00	2015-03-10 13:29:00	FA	FA	CN	FA	CN	CN	CN	FA	CN	FA	CN	FA	CN	CN	FA	FA	FA	FA	CN	CN	10	10	0
2016-07-22 06:00:00	2016-07-23 19:00:00	CN	CN	CN	FA	CN	CN	CN	CN	CN	FA	FA	FA	FA	CN	CN	CN	CN	FA	FA	CN	13	7	0
2021-10-31 01:57:00	2021-11-01 13:00:00	CN	CN	CN	CN	CN	CN	CN	CN	CN	CN	CN	CN	CN	CN	CN	CN	CN	FA	CN	CN	19	1	0
2021-11-01 13:00:00	2021-11-02 16:03:00	CN	CN	CN	CN	FA	CN	FA	FA	FA	FA	FA	CN	CN	FA	FA	CN	FA	FA	CN	CN	10	10	0
2022-01-17 18:01:00	2022-01-19 07:01:00	CN	CN	CN	CN	CN	CN	CN	CN	CN	FA	FA	CN	CN	CN	FA	CN	FA	FA	CN	FA	14	6	0
2022-04-16 04:17:00	2022-04-17 17:17:00	CN	CN	FA	FA	CN	CN	CN	CN	CN	FA	CN	FA	CN	CN	FA	CN	CN	FA	FA	CN	13	7	0
2022-04-19 04:41:00	2022-04-20 17:41:00	FA	FA	CN	FA	FA	CN	CN	CN	CN	FA	FA	FA	FA	CN	CN	CN	No Data	FA	FA	CN	9	10	1
2022-04-28 08:15:00	2022-04-29 21:15:00	CN	CN	CN	FA	CN	CN	CN	CN	CN	FA	FA	CN	CN	FA	FA	FA	FA	CN	CN	CN	13	7	0
2022-05-24 19:12:00	2022-05-26 08:12:00	FA	FA	CN	FA	CN	CN	CN	FA	CN	FA	CN	CN	CN	CN	FA	CN	FA	CN	CN	FA	12	8	0
2022-08-16 14:26:00	2022-08-18 00:00:00	FA	FA	No Data	No Data	CN	CN	CN	FA	CN	FA	CN	CN	CN	CN	FA	CN	CN	FA	CN	CN	12	6	2
2022-08-18 00:00:00	2022-08-18 21:00:00	FA	FA	CN	FA	CN	CN	CN	FA	CN	FA	CN	CN	CN	CN	FA	FA	FA	FA	CN	CN	11	9	0
2022-08-18 21:00:00	2022-08-19 18:14:00	FA	CN	CN	CN	CN	CN	CN	CN	CN	CN	CN	CN	CN	FA	FA	CN	FA	FA	CN	CN	15	5	0
2022-08-28 17:15:00	2022-08-30 00:00:00	FA	CN	CN	CN	CN	CN	No Data	CN	CN	FA	CN	CN	CN	CN	CN	CN	CN	FA	FA	CN	15	4	1
2022-08-30 00:00:00	2022-08-31 08:05:00	FA	CN	CN	CN	FA	CN	FA	FA	CN	FA	CN	CN	CN	CN	CN	CN	CN	FA	FA	CN	13	7	0
2022-11-30 08:04:00	2022-12-01 21:04:00	CN	CN	CN	FA	CN	CN	CN	CN	CN	CN	CN	CN	CN	FA	FA	FA	FA	CN	CN	FA	14	6	0
2023-03-03 16:19:00	2023-03-05 04:00:00	FA	CN	FA	FA	FA	CN	FA	FA	CN	FA	CN	FA	CN	CN	FA	CN	CN	FA	FA	CN	9	11	0
2023-03-05 04:00:00	2023-03-06 16:08:00	FA	CN	CN	FA	CN	CN	CN	FA	FA	FA	FA	FA	FA	CN	FA	CN	FA	FA	FA	CN	8	12	0
Total Correct Negative		9	16	27	12	22	29	24	17	28	9	24	18	27	26	13	25	15	7	18	25			
Total False Alarms		14	8	2	17	8	1	5	13	2	21	6	12	3	4	17	5	14	23	12	4			
Total No Data		7	6	1	1	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	1			

Figure 5.2: All Clear outcomes for SEPVAL challenge >10 MeV, 10 pfu non-event periods showing false alarms, correct negative, or no data for each model.

>100 MeV SEP Events	SPRINTS Post Eruptive 0-24 hrs	UMASEP-100	SAWS-ASPECS CME (SOHO)	SAWS-ASPECS CME (SOHO) electrons	SAWS-ASPECS flare + CME (SOHO)	SAWS-ASPECS flare + CME (SOHO) electrons	SAWS-ASPECS flare	SAWS-ASPECS flare electrons	SEPMOD	SEPSAT	SEPSTER (Parker Spiral)	SEPSTER2D CME	SFS-Update	ZEUS+iPATH CME	Total Hits	Total Misses	Total No Data
2011-06-07 07:20:00	Miss	Hit	Hit	Hit	Hit	Hit	Miss	Miss	Miss	Miss	Miss	No Data	Miss	Hit	6	7	1
2011-08-04 05:10:00	Hit	Hit	Hit	Hit	Hit	Hit	Hit	Hit	Hit	Miss	Hit	Hit	Miss	Hit	12	2	0
2011-08-09 08:25:00	Hit	Hit	Hit	Hit	Hit	Hit	Hit	Hit	Miss	Miss	No Data	Miss	Hit	No Data	9	3	2
2012-01-23 04:45:00	Hit	Hit	Hit	Hit	Hit	Hit	Hit	Hit	Hit	Miss	Hit	Hit	Miss	Hit	12	2	0
2012-01-27 19:00:00	Hit	Hit	Hit	No Data	Hit	No Data	Hit	No Data	Hit	Miss	Hit	Hit	Hit	Hit	10	1	3
2012-03-07 04:05:00	Miss	Hit	Hit	Hit	Hit	Hit	Hit	Hit	Miss	Miss	Miss	Hit	Hit	Miss	9	5	0
2012-03-13 18:10:00	Hit	Hit	Hit	Hit	Hit	Hit	Hit	Hit	Hit	Miss	Hit	Hit	Hit	Hit	13	1	0
2012-05-17 02:00:00	Miss	Hit	Hit	Hit	Hit	Hit	Hit	Hit	Miss	Miss	Miss	Miss	Miss	Hit	8	6	0
2013-04-11 09:40:00	Miss	Hit	Miss	Miss	Hit	Miss	Hit	Miss	Miss	Miss	Miss	Miss	Miss	Miss	3	11	0
2013-05-22 14:35:00	Hit	Hit	Hit	Hit	Hit	Hit	Hit	Hit	Hit	Miss	Hit	Hit	Hit	Miss	12	2	0
2014-01-06 08:30:00	No Data	Miss	Hit	Hit	No Data	No Data	No Data	No Data	Miss	No Data	Miss	Miss	No Data	Miss	2	5	7
2014-01-07 20:30:00	Hit	Hit	Hit	Hit	Hit	Hit	Hit	Hit	Miss	Miss	Hit	Hit	Hit	Hit	12	2	0
2015-10-29 04:35:00	No Data	Hit	Miss	Miss	No Data	No Data	No Data	No Data	Miss	Miss	Miss	Miss	No Data	No Data	1	6	7
2017-09-10 16:25:00	Hit	Hit	Hit	Hit	Hit	Hit	Hit	Hit	Hit	Miss	Hit	Hit	Hit	Hit	13	1	0
2021-10-28 16:35:00	Hit	Hit	Hit	Hit	Hit	Hit	Hit	Hit	Miss	Miss	Miss	Miss	Miss	Miss	8	6	0
2022-01-20 07:45:00	Miss	Hit	Miss	Miss	Hit	Hit	Hit	Hit	Miss	Miss	Miss	Miss	Miss	Miss	5	9	0
2022-03-28 12:45:00	Miss	Hit	Miss	Miss	Miss	Miss	Miss	Miss	Miss	Miss	Miss	Miss	Miss	Miss	1	13	0
Total Hits	9	16	13	12	14	12	13	11	6	0	7	8	7	8			
Total Misses	6	1	4	4	1	2	2	3	11	16	9	8	8	7			
Total No Data	2	0	0	1	2	3	2	3	0	1	1	1	2	2			

Figure 5.3: All Clear outcomes for SEPVAL challenge >100 MeV, 1 pfu SEP events showing hit, miss, or no data for each model.

>100 MeV Non-Event Start	>100 MeV Non-Event End	SPRINTS Post Eruptive 0-24 hrs	UMASEP-100	SAWS-ASPECS CME (SOHO)	SAWS-ASPECS CME (SOHO) electrons	SAWS-ASPECS flare + CME (SOHO)	SAWS-ASPECS flare + CME (SOHO) electrons	SAWS-ASPECS flare	SAWS-ASPECS flare electrons	SEPMOD	SEPSAT	SEPSTER (Parker Spiral)	SEPSTER2D CME	SFS-Update	ZEUS+iPATH CME	Total Correct Negatives	Total False Alarms	Total No Data
2011-03-07 01:05:00	2011-03-08 04:55:00	FA	CN	FA	FA	FA	FA	CN	CN	No Data	CN	FA	FA	CN	FA	5	8	1
2011-05-08 21:42:00	2011-05-10 10:42:00	CN	CN	CN	CN	CN	CN	CN	CN	CN	CN	CN	CN	CN	CN	14	0	0
2012-03-03 11:29:00	2012-03-04 19:00:00	CN	CN	FA	CN	FA	CN	CN	CN	CN	CN	CN	CN	CN	CN	12	2	0
2012-03-04 19:00:00	2012-03-05 17:30:00	CN	CN	FA	CN	FA	CN	FA	CN	CN	CN	CN	CN	FA	CN	10	4	0
2012-06-12 12:29:00	2012-06-14 01:29:00	CN	CN	CN	CN	CN	CN	CN	CN	CN	CN	CN	CN	CN	CN	14	0	0
2012-06-28 10:13:00	2012-06-29 23:13:00	CN	CN	CN	CN	CN	CN	CN	CN	CN	CN	CN	CN	CN	CN	14	0	0
2012-07-06 04:00:00	2012-07-07 07:45:00	FA	CN	FA	FA	FA	FA	FA	FA	CN	CN	CN	CN	CN	CN	7	7	0
2012-07-11 18:35:00	2012-07-12 22:15:00	FA	CN	CN	CN	FA	FA	FA	FA	CN	CN	CN	CN	FA	CN	8	6	0
2012-07-22 15:45:00	2012-07-23 13:45:00	No Data	CN	FA	FA	No Data	No Data	No Data	No Data	CN	CN	FA	FA	No Data	FA	3	5	6
2012-09-27 03:00:00	2012-09-28 04:45:00	CN	CN	CN	CN	CN	CN	CN	CN	FA	CN	CN	CN	CN	CN	13	1	0
2013-06-06 23:32:00	2013-06-08 12:32:00	CN	CN	CN	CN	FA	CN	FA	CN	CN	CN	CN	CN	CN	CN	12	2	0
2013-06-27 02:36:00	2013-06-28 15:36:00	CN	CN	CN	CN	CN	CN	CN	CN	CN	CN	CN	CN	CN	CN	14	0	0
2013-09-29 05:05:00	2013-09-30 17:00:00	CN	No Data	FA	FA	CN	CN	CN	CN	No Data	CN	CN	CN	CN	CN	10	2	2
2014-02-24 14:10:00	2014-02-25 22:10:00	CN	FA	FA	FA	FA	FA	FA	FA	CN	CN	CN	CN	FA	CN	6	8	0
2014-04-17 15:25:00	2014-04-18 19:55:00	FA	CN	FA	FA	FA	FA	FA	FA	CN	CN	CN	CN	CN	CN	7	7	0
2014-07-31 19:00:00	2014-08-02 08:00:00	CN	CN	CN	CN	CN	CN	CN	CN	CN	CN	CN	CN	CN	No Data	13	0	1
2014-09-10 02:55:00	2014-09-11 04:55:00	FA	FA	FA	FA	FA	FA	FA	FA	No Data	CN	CN	CN	FA	FA	3	10	1
2014-10-23 08:37:00	2014-10-24 21:37:00	CN	CN	CN	CN	CN	CN	CN	CN	CN	CN	CN	CN	CN	CN	14	0	0
2014-11-05 04:32:00	2014-11-06 17:32:00	CN	CN	CN	CN	CN	CN	CN	CN	CN	CN	CN	CN	CN	CN	14	0	0
2014-11-06 17:53:00	2014-11-08 06:53:00	CN	CN	CN	CN	CN	CN	FA	CN	CN	CN	CN	CN	FA	CN	12	2	0
2014-12-16 05:25:00	2014-12-17 18:25:00	CN	CN	CN	CN	FA	CN	FA	CN	CN	CN	CN	CN	CN	CN	12	2	0
2014-12-17 22:41:00	2014-12-19 11:41:00	CN	CN	FA	CN	FA	CN	FA	CN	CN	CN	CN	CN	CN	CN	11	3	0
2015-03-09 00:00:00	2015-03-10 13:29:00	CN	CN	CN	CN	FA	CN	CN	CN	CN	CN	CN	FA	CN	CN	12	2	0
2016-07-22 06:00:00	2016-07-23 19:00:00	CN	CN	CN	CN	FA	FA	FA	FA	CN	CN	CN	CN	CN	CN	10	4	0
2017-07-13 09:00:00	2017-07-14 14:45:00	FA	CN	FA	FA	FA	FA	CN	CN	CN	CN	CN	No Data	CN	CN	8	5	1
2017-09-04 00:40:00	2017-09-05 07:10:00	CN	CN	FA	FA	FA	FA	CN	CN	No Data	CN	CN	CN	CN	CN	9	4	1
2017-09-05 12:55:00	2017-09-06 16:30:00	FA	CN	FA	FA	FA	FA	FA	FA	FA	CN	CN	FA	FA	No Data	3	10	1
2021-05-28 03:00:00	2021-05-29 03:20:00	CN	CN	CN	CN	CN	CN	CN	CN	CN	CN	CN	CN	CN	FA	13	1	0
2021-10-31 01:57:00	2021-11-01 13:00:00	CN	CN	CN	CN	CN	CN	CN	CN	CN	CN	CN	CN	CN	CN	14	0	0
2021-11-01 13:00:00	2021-11-02 16:03:00	FA	CN	FA	FA	CN	CN	CN	CN	CN	CN	CN	CN	CN	CN	11	3	0
2022-01-17 18:01:00	2022-01-19 07:01:00	CN	CN	CN	CN	CN	CN	CN	CN	CN	CN	CN	CN	CN	CN	14	0	0
2022-04-01 14:40:00	2022-04-02 16:00:00	FA	FA	FA	FA	FA	FA	CN	CN	FA	CN	CN	FA	CN	FA	5	9	0
2022-04-16 04:17:00	2022-04-17 17:17:00	CN	CN	CN	CN	CN	CN	FA	CN	CN	CN	CN	CN	CN	CN	13	1	0
2022-04-19 04:41:00	2022-04-20 17:41:00	FA	CN	CN	CN	FA	FA	FA	FA	CN	CN	CN	No Data	FA	CN	7	6	1
2022-04-28 08:15:00	2022-04-29 21:15:00	CN	CN	CN	CN	CN	CN	CN	CN	FA	CN	CN	CN	CN	CN	13	1	0
2022-05-24 19:12:00	2022-05-26 08:12:00	CN	CN	CN	CN	CN	CN	CN	CN	CN	CN	CN	CN	CN	CN	14	0	0
2022-08-16 14:26:00	2022-08-18 00:00:00	CN	CN	CN	CN	CN	CN	CN	CN	CN	CN	CN	CN	CN	CN	14	0	0
2022-08-18 00:00:00	2022-08-18 21:00:00	CN	CN	FA	CN	FA	CN	CN	CN	CN	CN	CN	CN	CN	CN	12	2	0
2022-08-18 21:00:00	2022-08-19 18:14:00	CN	CN	CN	CN	CN	CN	CN	CN	CN	CN	CN	CN	CN	CN	14	0	0
2022-08-26 11:55:00	2022-08-27 12:20:00	FA	CN	FA	FA	FA	FA	FA	FA	CN	CN	CN	CN	CN	FA	6	8	0
2022-08-28 17:15:00	2022-08-30 00:00:00	CN	CN	CN	CN	FA	CN	CN	CN	CN	CN	CN	CN	CN	CN	13	1	0
2022-08-30 00:00:00	2022-08-31 08:05:00	FA	CN	FA	CN	FA	CN	CN	CN	CN	CN	CN	CN	CN	CN	11	3	0
2022-11-30 08:04:00	2022-12-01 21:04:00	CN	CN	CN	CN	CN	CN	CN	CN	CN	CN	CN	CN	CN	FA	13	1	0
2023-02-24 21:10:00	2023-02-26 02:50:00	FA	FA	FA	FA	FA	FA	FA	FA	CN	CN	CN	CN	FA	CN	5	9	0
2023-03-03 16:19:00	2023-03-05 04:00:00	FA	CN	CN	CN	FA	CN	FA	CN	CN	CN	CN	CN	CN	CN	11	3	0
2023-03-05 04:00:00	2023-03-06 16:08:00	CN	CN	CN	CN	FA	FA	FA	FA	CN	CN	CN	CN	CN	CN	10	4	0
Total False Alarms		14	4	19	14	25	15	18	11	4	0	2	5	8	7			
Total Correct Negatives		31	41	27	32	20	30	27	34	38	46	44	39	37	37			
Total No Data		1	1	0	0	1	1	1	1	4	0	0	2	1	2			

Figure 5.4: All Clear outcomes for SEPVAL challenge >100 MeV, 1 pfu non-event periods showing false alarms, correct negative, or no data for each model.

SEPVAL Median All Clear Scores	SEPVAL post-eruptive >10 MeV	SEPVAL post-eruptive >100 MeV	SEPVAL pre-eruptive >10 MeV	SEPVAL pre-eruptive >100 MeV
Percent Correct	0.73	0.74	0.55	-
Hit Rate	0.82	0.68	0.58	-
False Alarm Rate	0.23	0.21	0.46	-
False Alarm Ratio	0.25	0.53	0.36	-
Bias	1.02	1.35	0.91	-
Threat Score	0.56	0.37	0.42	-
HSS	0.47	0.36	0.07	-
TSS	0.47	0.38	0.07	-

Table 5.2: Median metrics for the SEPVAL challenge set and participating models. The models that contribute to the median scores are listed in Table 5.1.

For >10 MeV forecasts, the post-eruptive models in Table 5.2 have median scores of 82% Hit Rate, 23% False Alarm Rate, 0.56 Threat Score, and HSS and TSS of 0.47. The HSS and TSS mathematically have the same value for balanced data sets. The median Bias of 1.02 indicates that half the models have a tendency towards false alarms (>1.0) while the other half have a tendency towards misses (<1.0) with most models scores close to 1.0. The box plots in Figure 5.5 show a clear separation between Hit Rate and False Alarm Rate, which would be roughly equal for a balanced data set if forecasts had random skill. The median scores for SEPVAL post-eruptive models demonstrate that these models can discriminate between eruptions that lead to SEP events and those that do not. Figure 5.5 shows wide-ranging scores around these median values, but most models demonstrate some forecasting skill.

Pre-eruptive models have much lower forecasting skill for >10 MeV. With a median Hit Rate of 58% and a False Alarm Rate of 46%, just over half of the SEP events are hit and nearly half of the non-event periods are false alarms. The HSS is 0.07, slightly greater than zero, indicating that the skill is above random guessing, but not by much. All SEPVAL challenge periods are associated with strong flares and these models clearly have a difficult time discerning which active regions will produce only flares and which will produce flares and SEP events.

For >100 MeV forecasts, the post-eruptive models have lower median scores of 68% Hit Rate, 22% False Alarm Rate, 0.27 Threat Score, HSS of 0.36 and TSS of 0.40. The >100 MeV data set has 17 SEP events and 46 non-event periods, so the imbalance results in different values for HSS and TSS. Despite a drop in scores, the post-eruptive models show some skill at discriminating between event and non-event periods for the >100 MeV SEPVAL dataset.

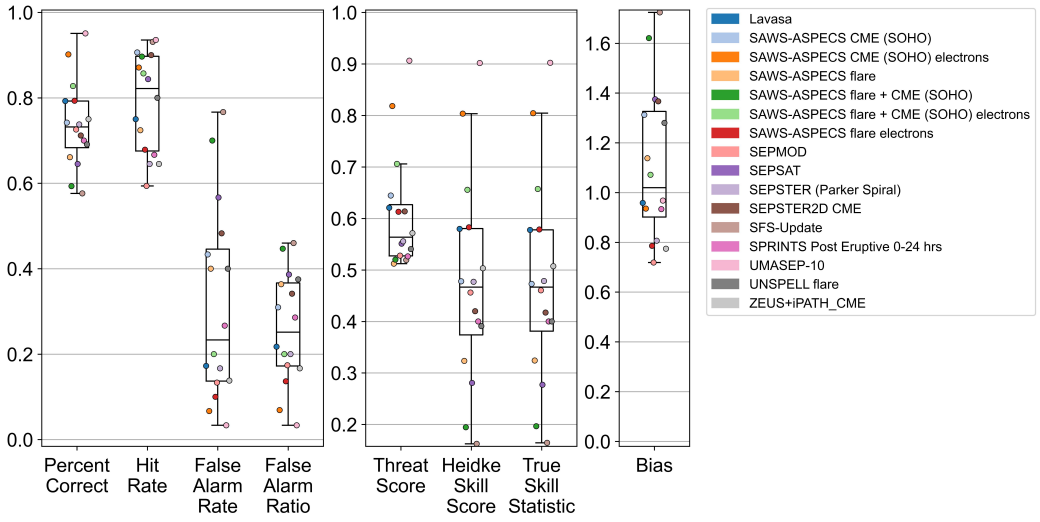


Figure 5.5: Summary box plots of SEPVAL All Clear metrics for post-eruptive forecasts associated with >10 MeV integral proton flux with 10 pfu flux threshold.

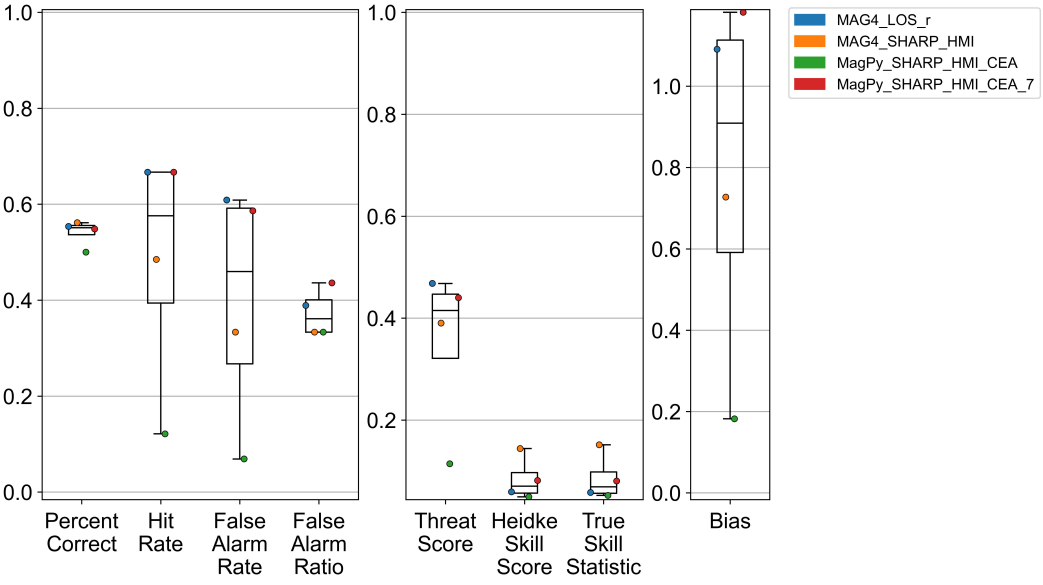


Figure 5.6: Summary box plots of SEPVAL All Clear metrics for pre-eruptive forecasts associated with >10 MeV integral proton flux with 10 pfu flux threshold.

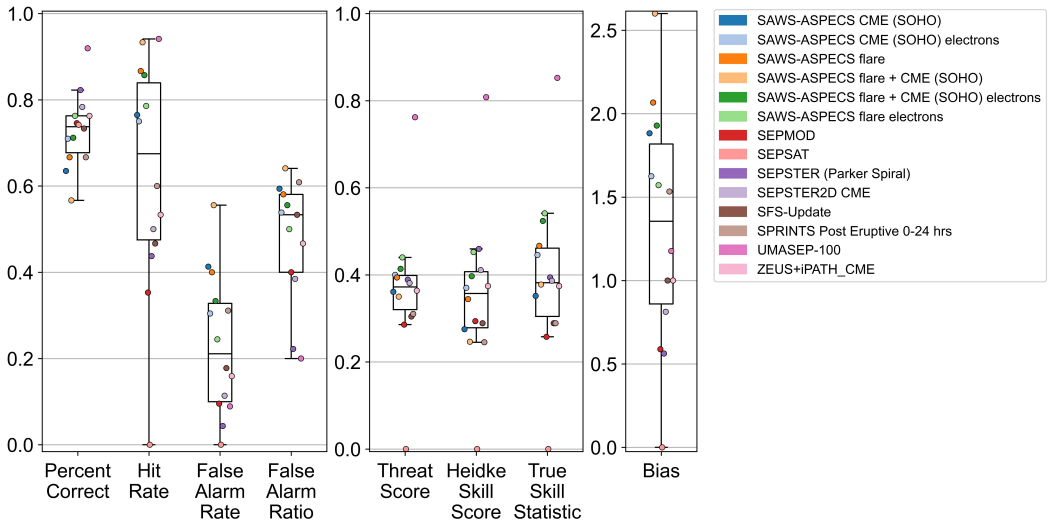


Figure 5.7: Summary box plots of SEPVAL All Clear metrics for post-eruptive forecasts associated with >100 MeV integral proton flux with 1 pfu flux threshold.

5.1.2 Probability

Many models participating in the SEPVAL effort provided forecasts for Probability of Occurrence, as indicated in Table 5.1. Most models were post-eruptive models, providing a likelihood that an SEP event would follow a particular flare or CME. MAG4 and MagPy predicted the probability that an SEP would occur in the next 24 hours based on the present magnetic conditions in active regions on the solar surface. Table 5.3 reports the median probability metrics for post-eruptive and pre-eruptive models and Figures 5.8–5.10 show the range of scores per model. The Brier score is the root-mean squared error of the predicted probability compared to the observed probability, thus a score of zero is the perfect score. To achieve a low Brier score, forecasted probabilities for SEP events should be high, close to 1.0, while forecasted probabilities for non-events should be low, close to zero. Figures 5.11–5.15 show the distribution of probabilities issued for SEP events and for non-event periods. A difference between these distributions indicates that a model is able to discriminate between the two cases. The Brier Skill Score compares model skill to climatology. In this study, the Brier Skill Score was calculated using the climatology published in Bain et al. (2021) as a reference, representing the average likelihood of an SEP event for any given day in Solar Cycle 24. This is an appropriate climatological comparison for pre-eruptive models like MAG4 and MagPy, but it is not an appropriate climatological reference for models triggered by flares and CMEs and is therefore left out of Table 5.3 for post-eruptive models. The Area Under the Curve (AUC) refers to the curve created in the ROC diagram, described in Section 3.3.2. An area of 0.5 represents random chance, less than 0.5 represents worse skill than random chance, and a value greater than 0.5 represents some skill with 1.0 being a perfect score. Figures 5.16, 5.17, and 5.18 show the ROC curves for SEPVAL.

SEPVAL Median	SEPVAL post-eruptive		SEPVAL pre-eruptive	
Probability Scores	>10 MeV	>100 MeV	>10 MeV	>100 MeV
Brier Score	0.23	0.17	0.22	-
Brier Skill Score	-	-	−0.003	-
Area Under the Curve	0.76	0.78	0.56	-

Table 5.3: Median metrics for the SEPVAL challenge set and participating models.

The post-eruptive models reported in Table 5.3 and in Figure 5.8 achieve a median Brier score and individual scores that are fairly low (i.e., good) for both >10 and >100 MeV. This indicates that, across models, forecast probabilities for SEP events must typically be on the higher side while those for non-events are lower. The pre-eruptive models have a similar median Brier Score for >10 MeV, however the Brier Skill Score is −0.003, indicating no skill compared to using the average probability of SEP occurrence for Solar Cycle 24 as a prediction.

Figures 5.11 – 5.15 help to further interpret the Brier score. These raincloud plots group forecasts associated with observed SEP events (top) and forecasts associated with non-event periods (bottom) for >10 MeV (left) and >100 MeV (right).

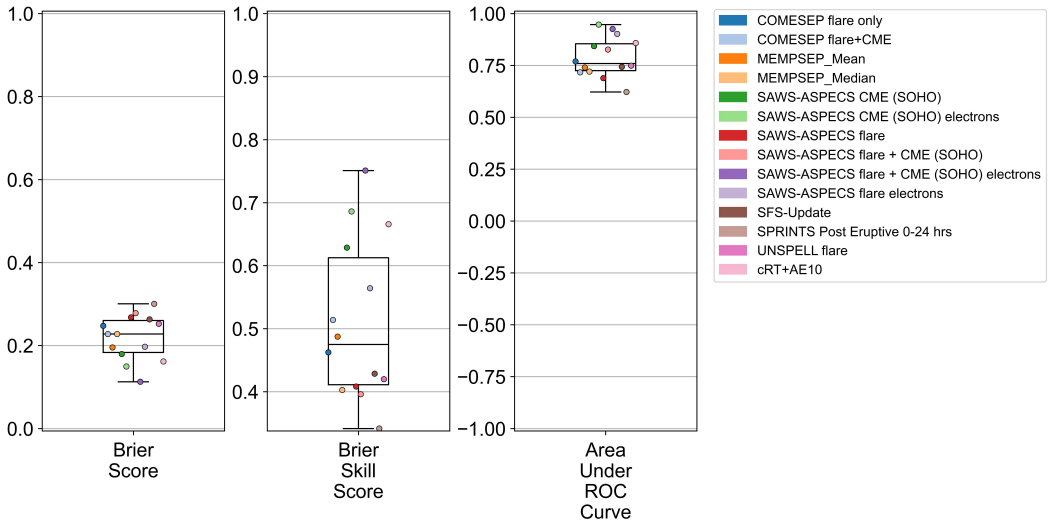


Figure 5.8: Summary box plots of SEPVAL Probability metrics for post-eruptive model forecasts associated with >10 MeV integral proton flux with 10 pfu flux threshold.

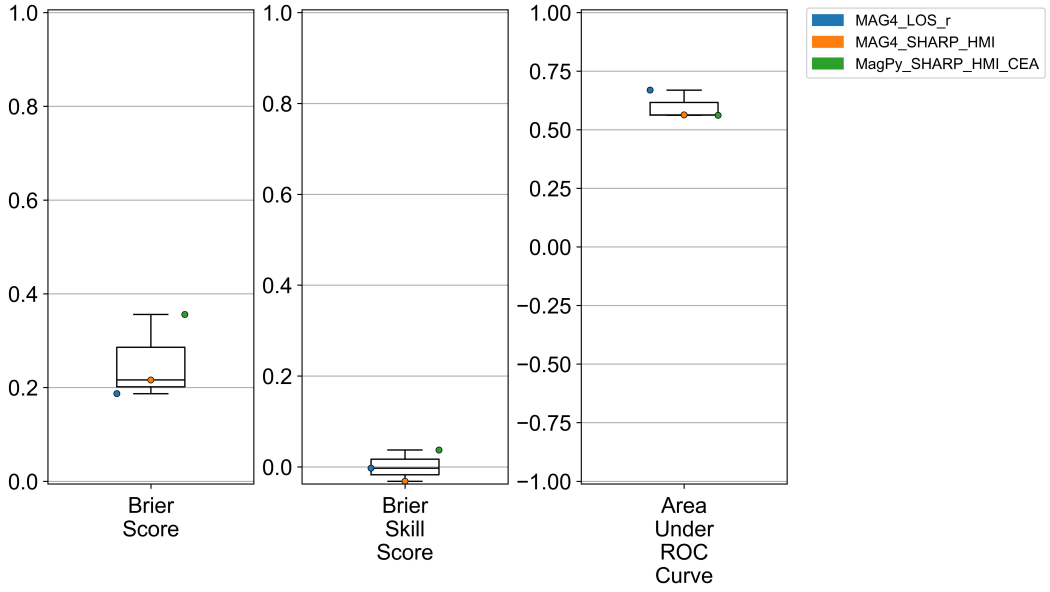


Figure 5.9: Summary box plots of SEPVAL Probability metrics for pre-eruptive model forecasts associated with >10 MeV integral proton flux with 10 pfu flux threshold.

The points show the forecasted probability values and the shaded region shows the kernel density estimator (KDE) of the distribution of probabilities. Ideally, SEP and No SEP plots should have different distributions that allow the model to discriminate between the two cases. Figure 5.11 shows the SPRINTS probability

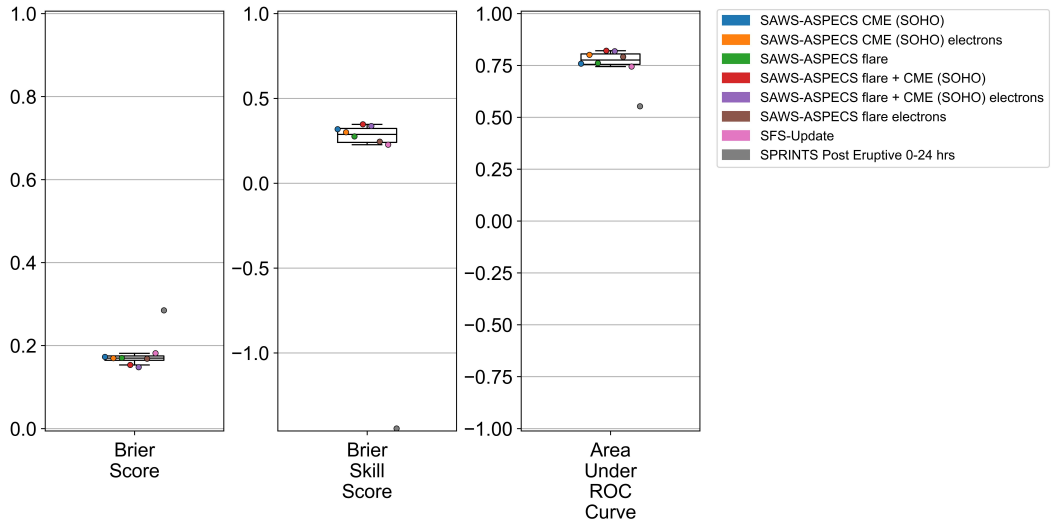


Figure 5.10: Summary box plots of SEPVAL Probability metrics for post-eruptive model forecasts associated with >100 MeV integral proton flux with 1 pfu flux threshold.

forecasts. SPRINTS issues a probability forecast every minute while a flare is ongoing. The figure shows that forecasts associated with >10 MeV SEPs achieve higher probabilies than those associated with non-events, indicating good discrimination. The >100 MeV case is clearly more challenging, nonetheless the two distributions are different. Figure 5.12 shows raincloud plots for COMESEP, MEMPSEP, PPS (SFS-Update), UNSPELL, and cRT+AE10. Each of these models produced a single probability forecast per SEPVAL challenge period. As indicated by the Brier Scores, it is seen that most models have different probability distributions between SEP and non-event periods. The SAWS-ASPECS models in Figures 5.13 and 5.14 show an interesting comparison that clearly demonstrates how the use of electrons to suppress false alarm increases discrimination between events and non-events. Finally, Figure 5.15 shows that the pre-eruptive models do not produce probability distributions that discriminate well between the two types of challenge periods.

The median AUC for post-eruptive models is 0.76 and all of the individual scores are above 0.5, which indicates forecasting skill. The ROC curve reflects how the choice of a probability threshold to convert probability forecasts to binary All Clear forecasts affects the False Alarm Rate and the Hit Rate. It is desirable to have a curve that extends to the top left corner where False Alarm Rate is low but Hit Rate is high. The ROC curves for SAWS-ASPECS in Figure 5.16 show very good skill in both >10 and >100 MeV, particularly for the "electrons" model versions that use electron flux as input to reduce false alarms. Figure 5.17 shows that the other SEPVAL post-eruptive models have similar skill to the non-electron version of SAWS-ASPECS for >10 MeV. The cRT+AE10 model stands out in this group as a high performer in this metric. For >100 MeV, the SAWS-ASPECS models and PPS (SFS-Update) shows similar skill to the >10 MeV counterparts, as reflected in the

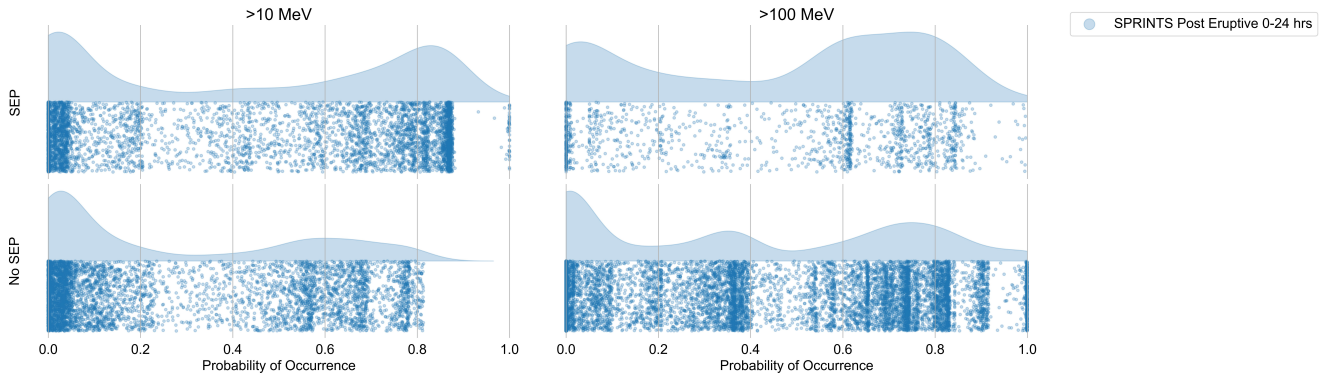


Figure 5.11: SEPVAL probability distributions for SPRINTS associated with >10 MeV integral proton flux with 10 pfu flux threshold and >100 MeV integral proton flux with 1 pfu flux threshold.

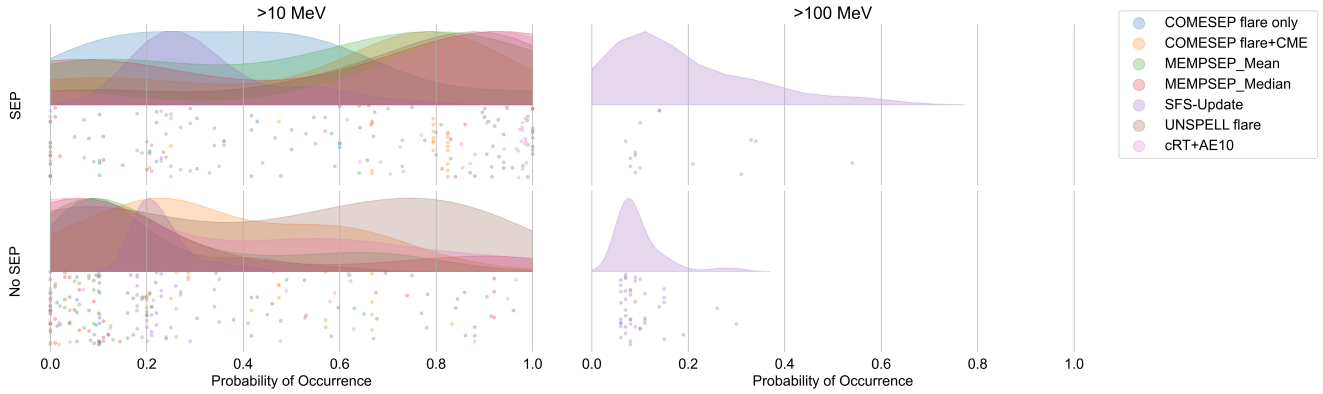


Figure 5.12: Probability distributions for SEPVAL post-eruptive models associated with >10 MeV integral proton flux with 10 pfu flux threshold and >100 MeV integral proton flux with 1 pfu flux threshold.

similar median AUC scores. The median AUC for pre-eruptive models is 0.56, just above random chance. The ROC curves in Figure 5.18 show performance of MAG4 and MagPy SHARP versions slightly above random guessing with MAG4_LOS_r demonstrating the most skill.

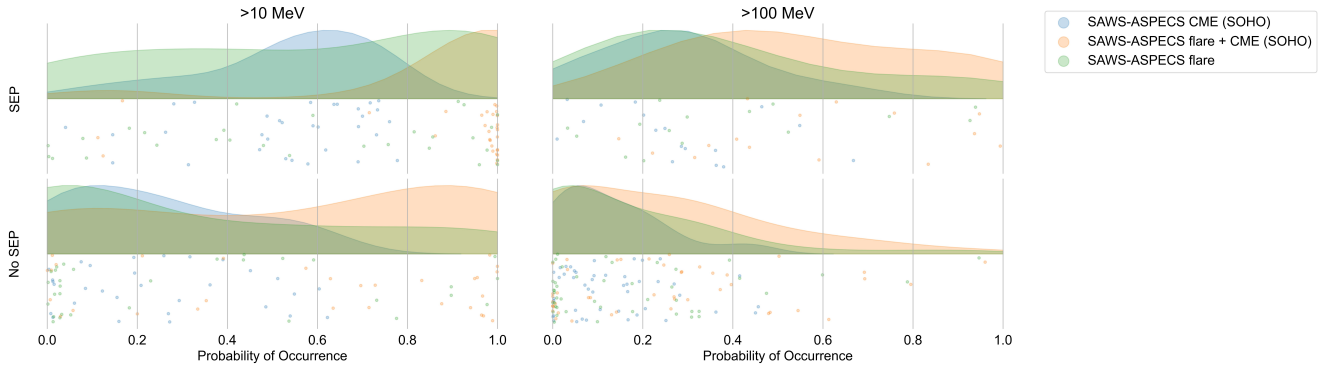


Figure 5.13: SEPVAL probability distributions for selected SAWS-ASPECS models associated with >10 MeV integral proton flux with 10 pfu flux threshold and >100 MeV integral proton flux with 1 pfu flux threshold.

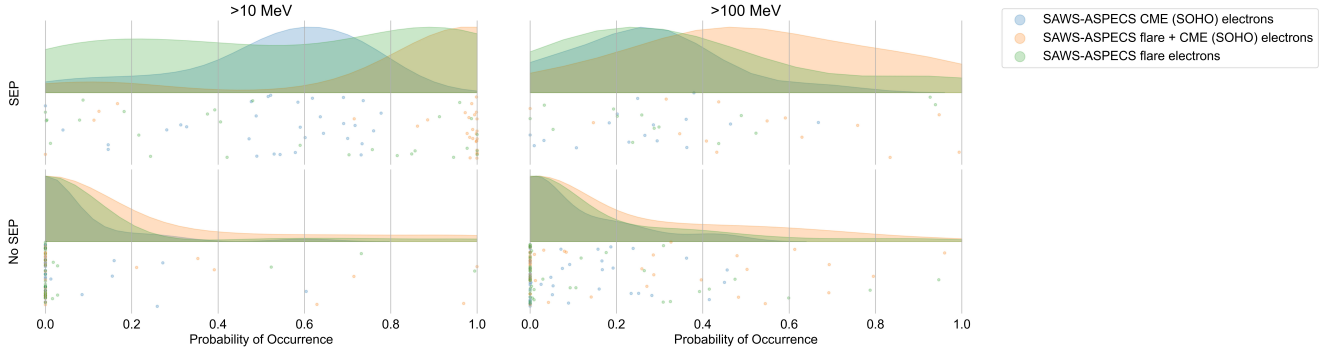


Figure 5.14: SEPVAL probability distributions for selected SAWS-ASPECS electrons models associated with >10 MeV integral proton flux with 10 pfu flux threshold and >100 MeV integral proton flux with 1 pfu flux threshold.

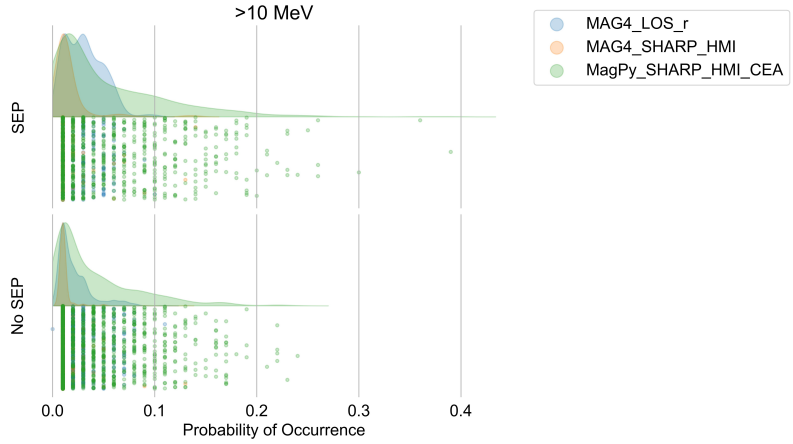


Figure 5.15: SEPVAL probability distributions for selected pre-eruptive models associated with >10 MeV integral proton flux with 10 pfu flux threshold.

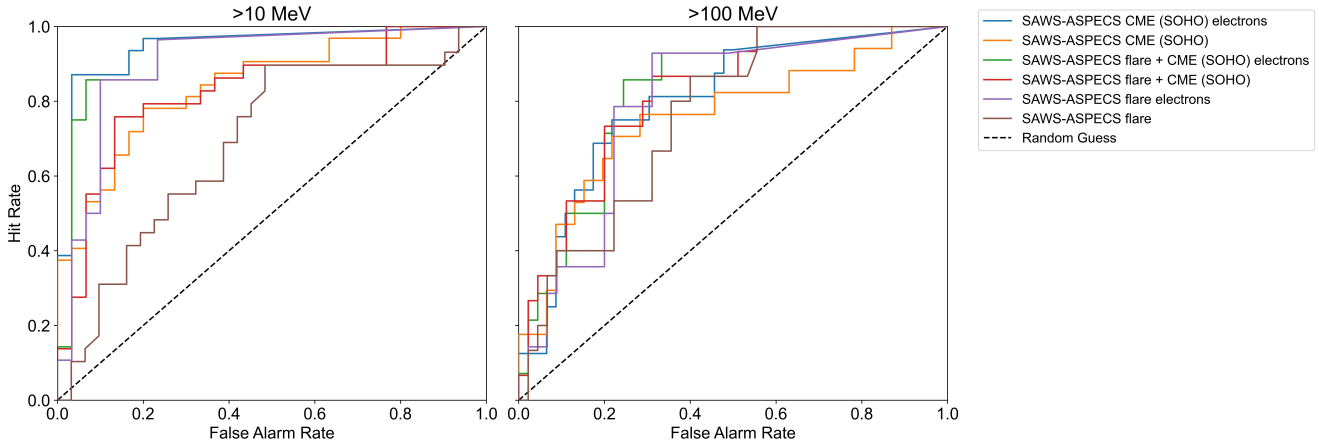


Figure 5.16: ROC curves for selected SEPVAL SAWS-ASPECS results associated with >10 MeV integral proton flux with 10 pfu flux threshold and >100 MeV integral proton flux with 1 pfu flux threshold.

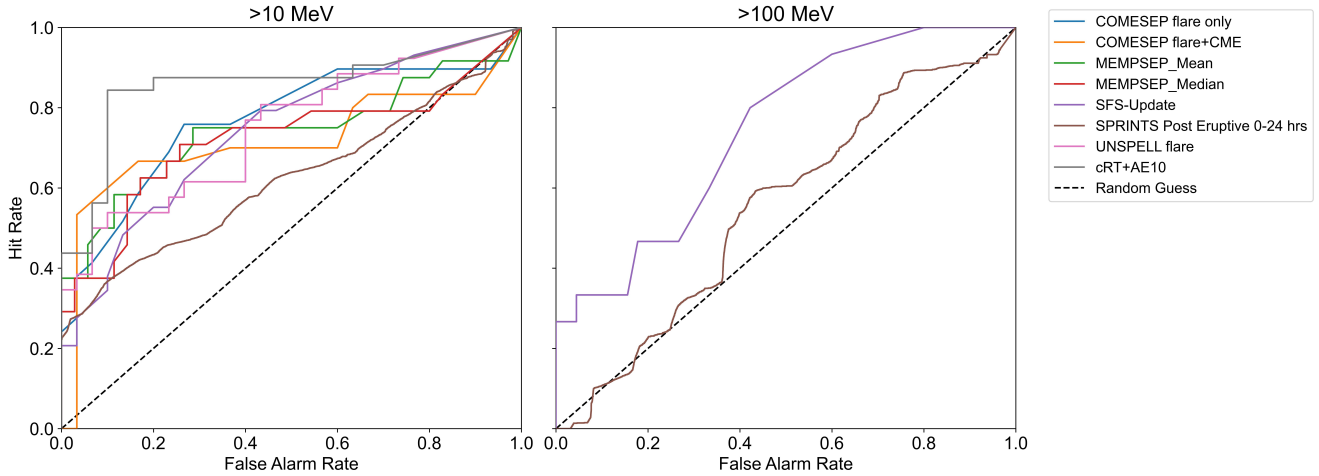


Figure 5.17: ROC curves for selected SEPVAL post-eruptive models associated with >10 MeV integral proton flux with 10 pfu flux threshold and >100 MeV integral proton flux with 1 pfu flux threshold.

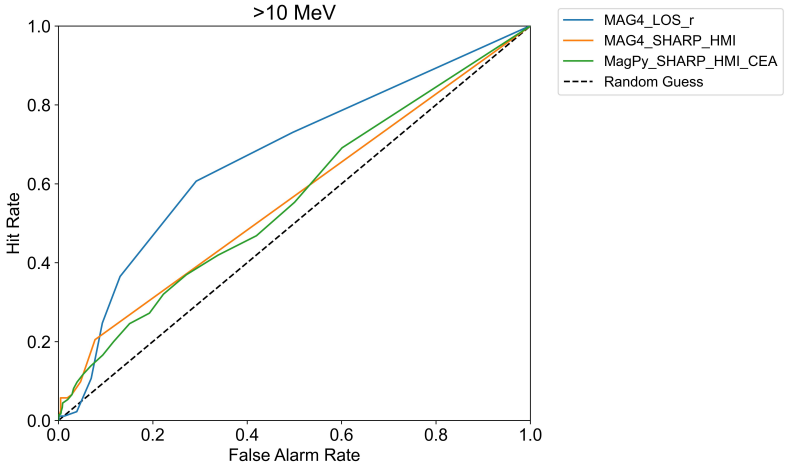


Figure 5.18: ROC curves for SEPVAL pre-eruptive models associated with >10 MeV integral proton flux with 10 pfu flux threshold.

5.1.3 Onset Peak and Maximum Flux

Many models aim to predict the peak flux of an impending SEP event. With SPHINX, we consider both the onset peak and maximum flux. Without knowing the details of each model, it is sometimes unclear which peak flux prediction a particular model intends to provide, so forecasts are evaluated against both.

SEPVAL Median	Onset Peak		Maximum Flux	
Flux Metrics	>10 MeV	>100 MeV	>10 MeV	>100 MeV
Median Log Error	0.06	−0.15	−0.15	−0.50
Median Absolute Log Error	0.94	0.89	0.86	0.96
Pearson Correlation Coefficient	0.37	0.33	0.39	0.34
Spearman Correlation Coefficient	0.29	0.22	0.39	0.24

Table 5.4: Median metrics for the SEPVAL challenge set and participating models. The models that contribute to the median scores are listed in Table 5.1. The Pearson Correlation Coefficient is calculated in log space. The onset peak refers to the initial rise of the SEP event while the Max Flux refers to the maximum flux value measured during the SEP event.

The median metrics for onset peak and maximum flux are reported in Table 5.4. Figures 5.19 – 5.26 show the spread of the metrics in box plots and the Log Error distribution for each model in combined histograms. The median log error for both the >10 MeV onset peak and the max flux are near to zero, indicating that the model predictions as a whole are centered on the observed peak values with little bias. The median absolute log error, however, indicates that the predictions are generally different from the observed fluxes by an order of magnitude. Both the median Pearson and Spearman correlation coefficients indicate that model predictions trend with observations, but not well. The box plots in Figures 5.19 and 5.21 show that the individual models have very wide-ranging errors and correlation values. The histograms in Figures 5.20 and 5.22 show that most SEPVAL models have a bias close to zero, however predictions can differ from observations by multiple orders of magnitude.

The median log error for >100 MeV onset peak and maximum flux are both negative (although the onset peak error is small) indicating that the models have a tendency to underpredict >100 MeV peak flux. The predictions also tend to differ from observed values by an order of magnitude or more, as indicated by the median absolute log error. The correlation coefficients are low. Figures 5.23 – 5.26 show that the metrics widely vary per model and the distribution of log errors have an overall bias to underpredict >100 MeV flux. Again, the errors range over a few orders of magnitude.

Many models produce onset peak and max flux forecasts within an order of magnitude of the observed value, but the potential for very large errors for individual forecasts indicates that more improvements need to be made to increase the reliability of peak flux forecasts.

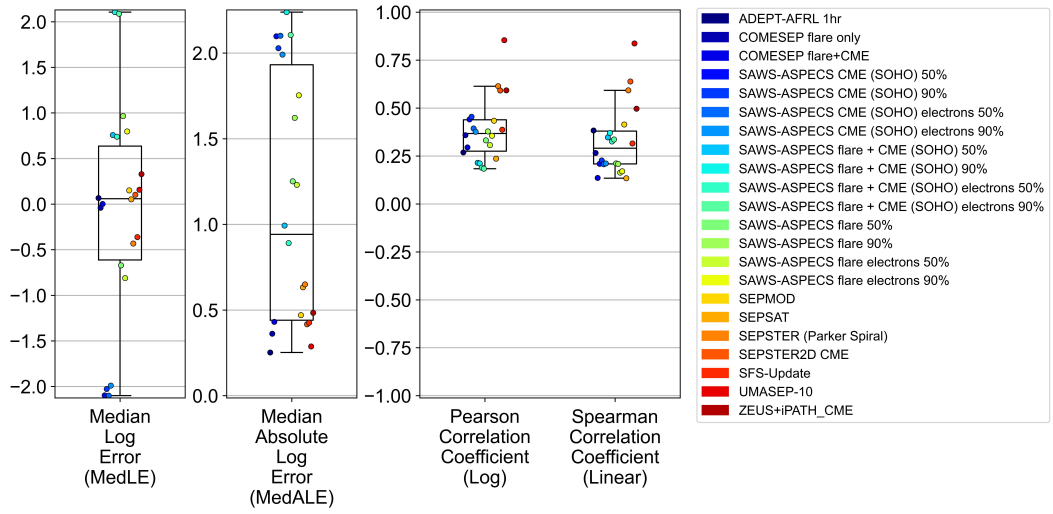


Figure 5.19: Summary box plots of SEPVAL onset peak metrics for forecasts associated with >10 MeV integral proton flux with 10 pfu flux threshold.

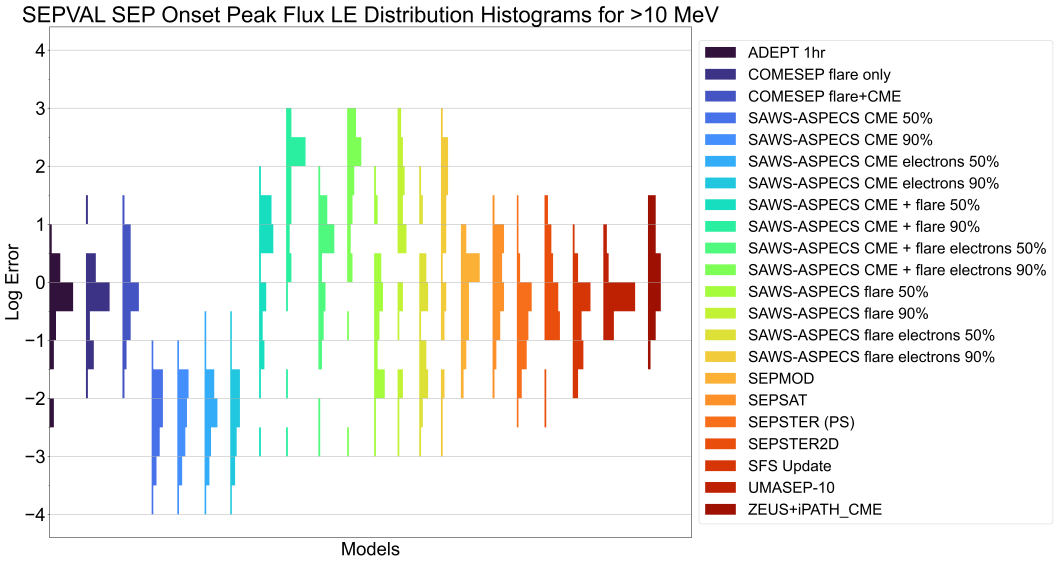


Figure 5.20: Summary histograms of SEPVAL onset peak Log Error for forecasts associated with >10 MeV integral proton flux with 10 pfu flux threshold.

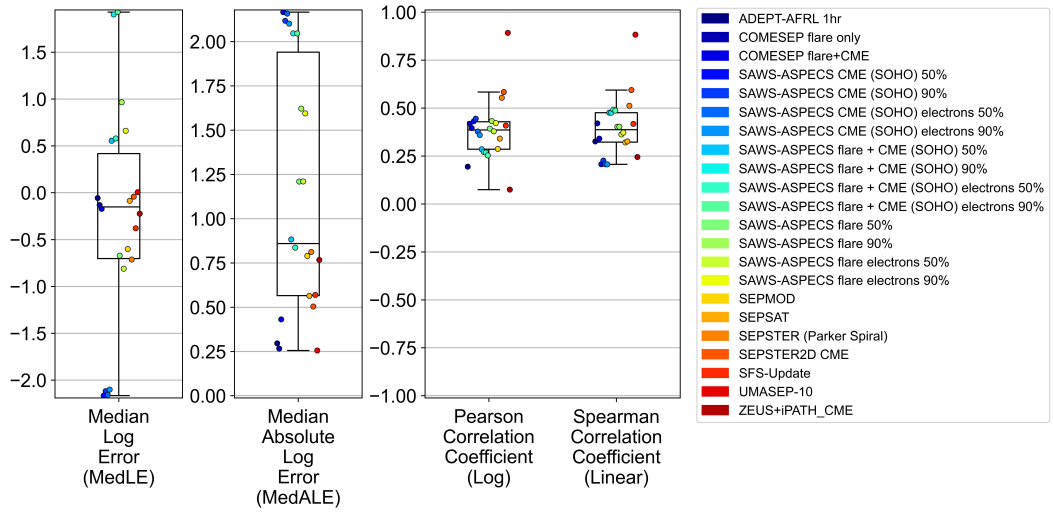


Figure 5.21: Summary box plots of SEPVAL max flux metrics for forecasts associated with >10 MeV integral proton flux with 10 pfu flux threshold.

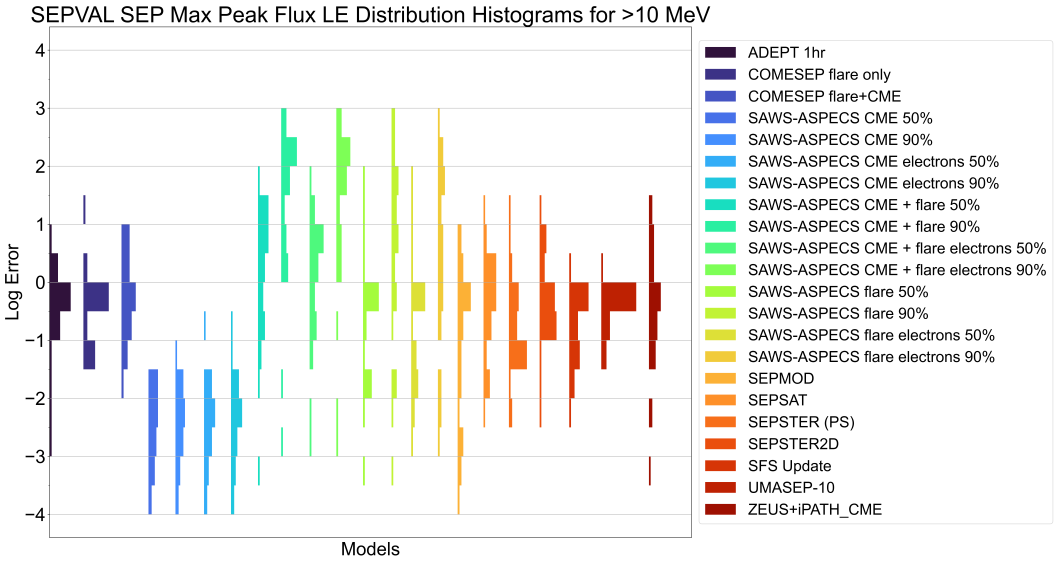


Figure 5.22: Summary histograms of SEPVAL maximum flux Log Error for forecasts associated with >10 MeV integral proton flux with 10 pfu flux threshold.

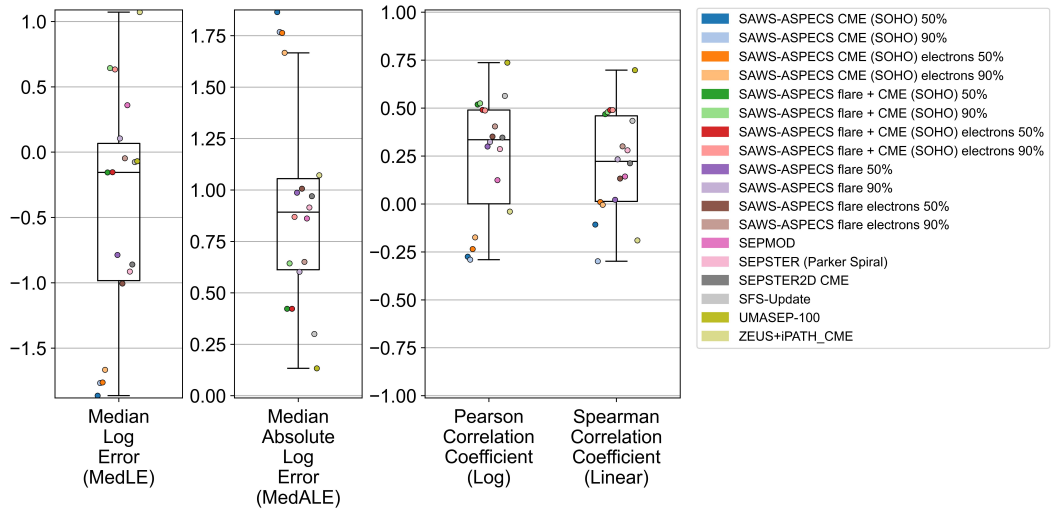


Figure 5.23: Summary box plots of SEPVAL onset peak metrics for forecasts associated with >100 MeV integral proton flux with 1 pfu flux threshold.

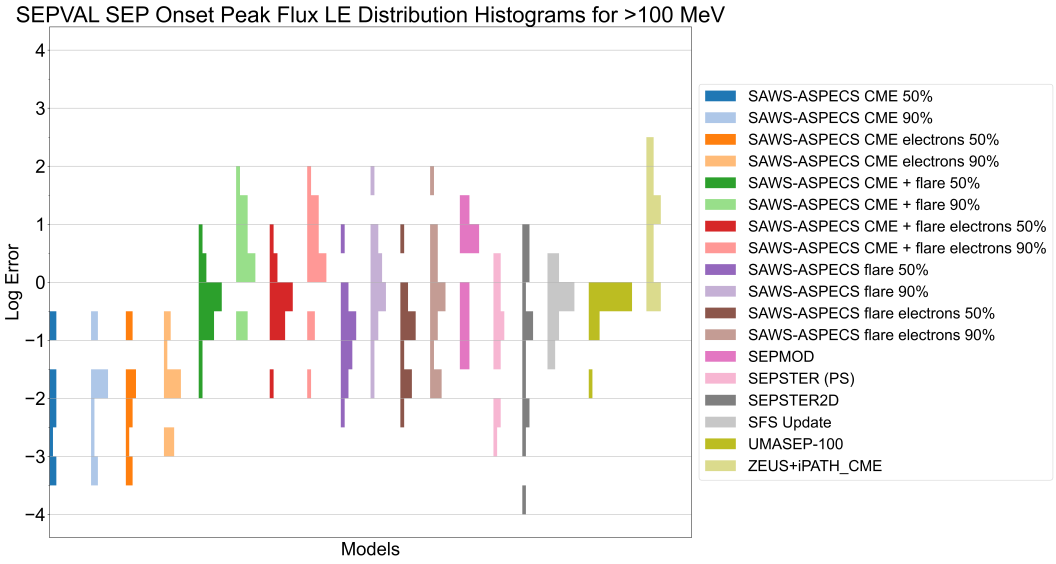


Figure 5.24: Summary histograms of SEPVAL onset peak Log Error for forecasts associated with >100 MeV integral proton flux with 1 pfu flux threshold.

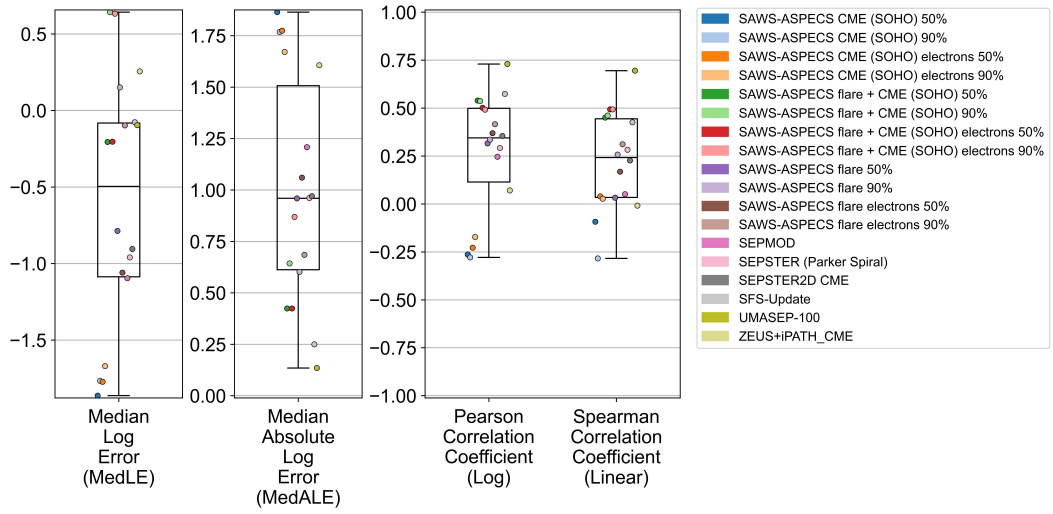


Figure 5.25: Summary box plots of SEPVAL max flux metrics for forecasts associated with >100 MeV integral proton flux with 1 pfu flux threshold.

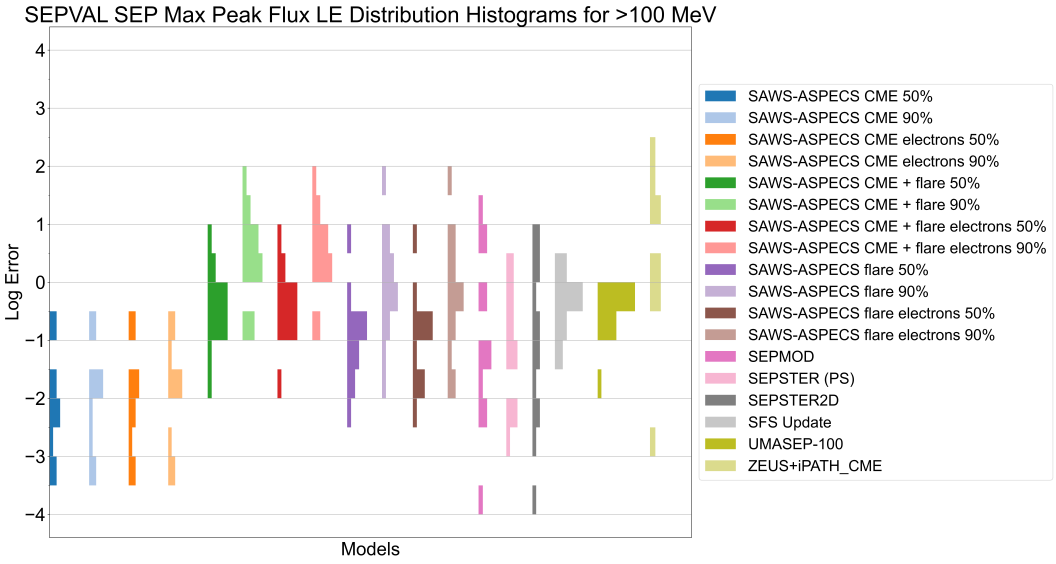


Figure 5.26: Summary histograms of SEPVAL maximum flux Log Error for forecasts associated with >100 MeV integral proton flux with 1 pfu flux threshold.

5.1.4 SEPVAL Group Results Summary

Here we summarize the main takeaways of the SEPVAL challenge from the perspective of the overall group of participating models. The SEPVAL validation effort can be viewed as an idealized real-time-like validation with high quality inputs and a balanced data set of a roughly equal number of SEP events and non-event periods. The All Clear and probability metrics demonstrate that the participating models are able to discriminate between the SEPVAL challenge SEP events and non-events. The flare and CME post-eruptive models are more successful than the pre-eruptive models in all cases. Performance is stronger for >10 MeV events than >100 MeV events, but the models demonstrate skill for both. Onset peak and max flux prediction median values show little bias, but the accuracy is low with a very wide range in errors reaching up to 3 orders of magnitude or more between predicted and observed peak flux values.

For >10 MeV, the post-eruptive models have a median Hit Rate of 82%, False Alarm Ratio of 25% and HSS of 0.47. The median Brier Score is a fairly good 0.23 and the median AUC is a fairly skillful 0.76. The ROC curves demonstrate significant skill over random guessing. The pre-eruptive models show less success in All Clear and probability forecasting, achieving metrics just slightly higher than random chance with the best performance from the MAG4 LOS approach. The onset peak and max flux predictions are generally centered around the observed fluxes, with a median log error close to zero, but there is a very large spread, giving a median value of ≈1 order of magnitude with errors reaching up to 3 or more orders of magnitude.

For >100 MeV, model performance is generally lower than for >10 MeV but models are able to discriminate between >100 MeV events and non-event periods. The median Hit Rate is 68%, the False Alarm Ratio increases to 53%, and the HSS drops to 0.36. The median Brier Score is good at 0.17 and the median AUC of 0.78 shows skill. The onset peak and max flux predictions have a bias towards underprediction with median values of -0.15 and -0.50, respectively. The absolute log error in peak flux predictions has a median value of ≈1 order of magnitude, with a spread in errors of up to more than 3 orders of magnitude.

5.2 SEP Scoreboard Group Results

The SEP Scoreboards have aggregated real-time forecasts from 10 different SEP forecasting models and SWPC for over 4 years. These forecasts represent the true model performance in a real-time setting and their utility for operations. As described in Section 4.2, these results include the performance of the full forecasting chain including real-time data-quality and gaps, model approach, model robustness and run time, model version updates, and human-in-the-loop support. This analysis can provide an assessment of AWT that reflects the availability of forecasts for use by operators. The underlying dataset is a true climatological sample of SEP event and non-event periods, the imbalance of which depends on each model's triggers and cadence. The performance reported here can inform SRAG about the use of models in operations as well as establish realistic target metrics that define the state of the art for real-time SEP model forecasting.

In this section we will discuss the group performance (median and distribution) for the SEP Scoreboards. Section 6 describes the performance of each individual model for the Scoreboards and SEPVAL (when applicable) in great detail. As was done for the SEPVAL group comparison, post-eruptive and pre-eruptive models are evaluated separately. Table 5.5 shows which models are included in the median metrics for All Clear, probability, onset peak and maximum flux. Most models are post-eruptive models triggered by flares or CMEs. MAG4 and MagPy forecasts are produced on a pre-eruptive basis using magnetograms as input. SWPC Day 1 forecasts are also considered to be in the pre-eruptive category while SWPC Warnings are in the post-eruptive category. Deoverlapped All Clear scores (see Section 3.5) are used for GSU, MAG4, MagPy, and UMASEP.

Scoreboard Model	Model Type	Included in Median Metrics
GSU All Clear	pre-eruptive	All Clear, Probability
MAG4_LOS_FEr	pre-eruptive	All Clear, Probability
MAG4_LOS_r	pre-eruptive	All Clear, Probability
MAG4_SHARP	pre-eruptive	All Clear, Probability
MAG4_SHARP_FE	pre-eruptive	All Clear, Probability
MAG4_SHARP_HMI	pre-eruptive	All Clear, Probability
MagPy_SHARP_HMI_CEA (19%)	pre-eruptive	All Clear, Probability
ENLIL+SEPMOD	post-eruptive	All Clear, Peak
SEPSTER (Parker Spiral)	post-eruptive	All Clear, Peak
SEPSTER (WSA-ENLIL)	post-eruptive	All Clear, Peak
SEPSTER2D	post-eruptive	All Clear, Peak
SPRINTS Post Eruptive 0-24 hrs	post-eruptive	All Clear, Probability
SWPC Warning	post-eruptive	All Clear
UMASEP-10, UMASEP-100	post-eruptive	All Clear, Peak
ZEUS+iPATH CME	post-eruptive	All Clear, Peak

Table 5.5: Models included in SEP Scoreboard median metrics for All Clear, Probability, and Peak (Onset Peak and Max Flux).

5.2.1 All Clear

All Clear performance is described through numerous metrics and skill scores, a subset of which are selected as a focus for this report. Some of these metrics and skill scores are impacted by the level of imbalance in the dataset used for validation, as discussed in Section 3.4. The SEPVAL events are an approximately balanced data set while the SEP Scoreboards are highly imbalanced between SEP and non-event periods, therefore the numerical values of many metrics cannot be directly compared between SEPVAL and the SEP Scoreboards. It is also a challenge to compare scores from different models across the Scoreboards due to their own individual time coverage and climatologies, however Ahmadzadeh et al. (2023) show that metrics become less sensitive to imbalance as the imbalance increases. Since the SEP Scoreboard models are all in the high imbalance regime, we assume that skill scores can be approximately compared across models.

Scoreboard Median	Scoreboard Post-eruptive		Scoreboard Pre-eruptive	
All Clear Scores	>10 MeV	>100 MeV	>10 MeV	>100 MeV
Percent Correct	0.95	0.99	0.55	-
Hit Rate	0.61	0.18	0.71	-
False Alarm Rate	0.04	0.006	0.46	-
False Alarm Ratio	0.67	0.90	0.96	-
Bias	1.80	2.24	16.9	-
Threat Score	0.26	0.06	0.04	-
HSS	0.28	0.09	0.03	-
TSS	0.50	0.14	0.21	-

Table 5.6: Median All Clear metrics for the SEP Scoreboard. The models that contribute to the median scores are listed in Table 5.5.

Selected median All Clear scores are reported in Table 5.6. For the post-eruptive models, the >10 MeV median Hit Rate is 61%, the False Alarm Rate is 4%, and the False Alarm Ratio is 67%. In the SEP Scoreboards, models are hitting only about half of the observed SEP with a range as low as ≈10% up to ≈90%+, shown in Figure 5.27. The human-driven SWPC Warning has the highest Hit Rate (although Warnings issued after SEP thresholds were crossed were not included in this analysis). The False Alarm Rate is much lower than SEPVAL, however due to the highly imbalanced climatology, even a small percentage of false alarms results in a number of false alarms that exceeds the number of observed SEP events. This is reflected in the False Alarm Ratio which reports that 67% of yes forecasts are false alarms. The median bias of 1.80 reflects this tendency towards false alarms while the Threat Score (0.26), HSS (0.28), and TSS (0.50) are fairly low. The False Alarm Ratio, Threat Score and HSS are more sensitive to the number of false alarms with respect to hits (False Alarm Ratio), resulting in very little skill. The TSS is more sensitive to the Hit Rate and False Alarm Rate, so the score is higher. This highlights the fact that, in a climatological scenario, the Hit Rate needs to be high while the False Alarm Rate needs to be reduced to extremely small values to achieve

high skill scores. Deoverlapped UMASEP-10 achieved the highest HSS = 0.46 with a False Alarm Rate of 3% and Hit Rate of 69%. SEPSTER (Parker Spiral) achieved the second highest HSS = 0.44 with a False Alarm Rate of 1.8% and a Hit Rate of 62%.

The >10 MeV All Clear scores for the pre-eruptive models are much lower. As seen in the SEPVAL results in Section 5.1, these types of models have a difficult time discerning between active periods that generate SEPs and those that do not. The median Hit Rate is 71%, but the False Alarm Rate is 46%, and the False Alarm Ratio is 95%. These models forecast an event will occur nearly half of the time when conditions are quiet and that nearly all of the yes forecasts are false alarms. The median Threat Score (0.04) and HSS (0.03) are just barely above zero, indicating no skill above random chance. Figure 5.28 shows that there are a range of Hit Rates and False Alarm Rates, but that the False Alarm Ratios are all very high and the skill scores are all low.

All Clear performance for >100 MeV events is poor for post-eruptive models (no pre-eruptive models predict >100 MeV). The range of scores is plotted in Figure 5.29. The low Hit Rates and high False Alarm Ratios result in very little forecasting skill, with the exception of UMASEP-100.

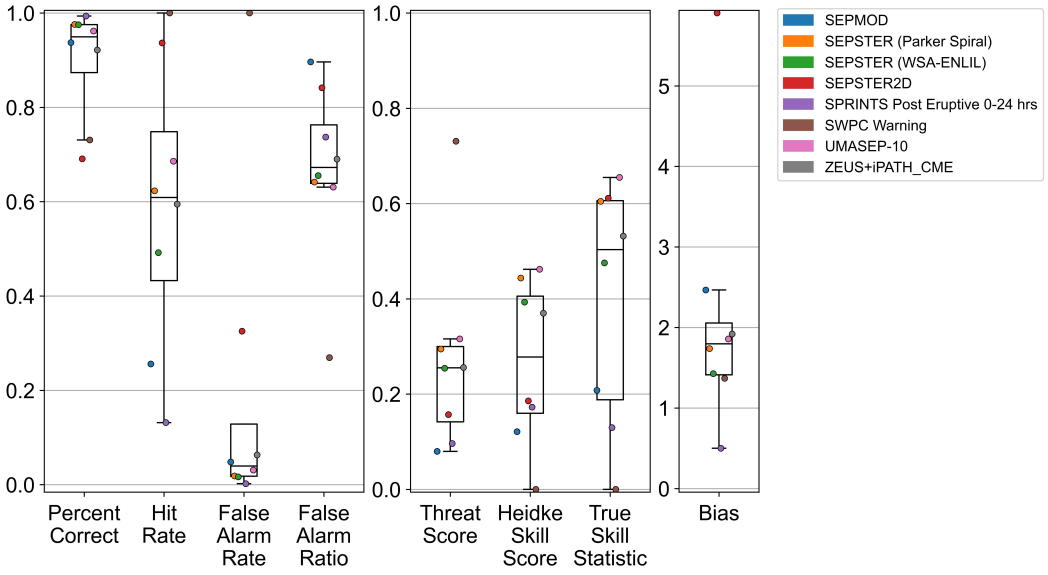


Figure 5.27: All Clear, Post-eruptive, >10 MeV: Summary box plots of SEP Scoreboard All Clear metrics for post-eruptive forecasts associated with >10 MeV integral proton flux with 10 pfu flux threshold.

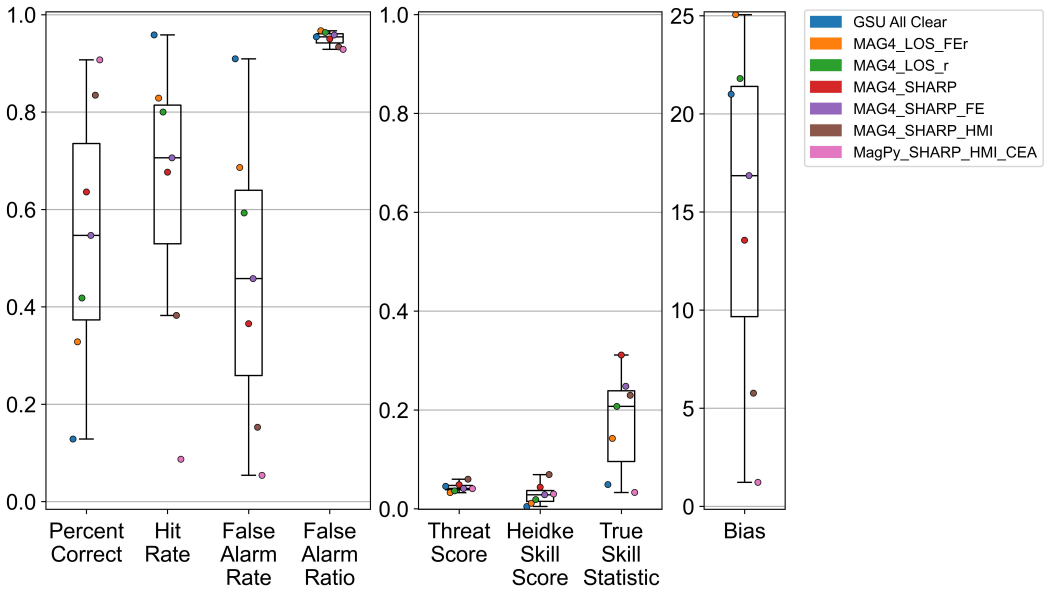


Figure 5.28: All Clear, Pre-eruptive, >10 MeV: Summary box plots of SEP Scoreboard All Clear metrics for pre-eruptive forecasts associated with >10 MeV integral proton flux with 10 pfu flux threshold.

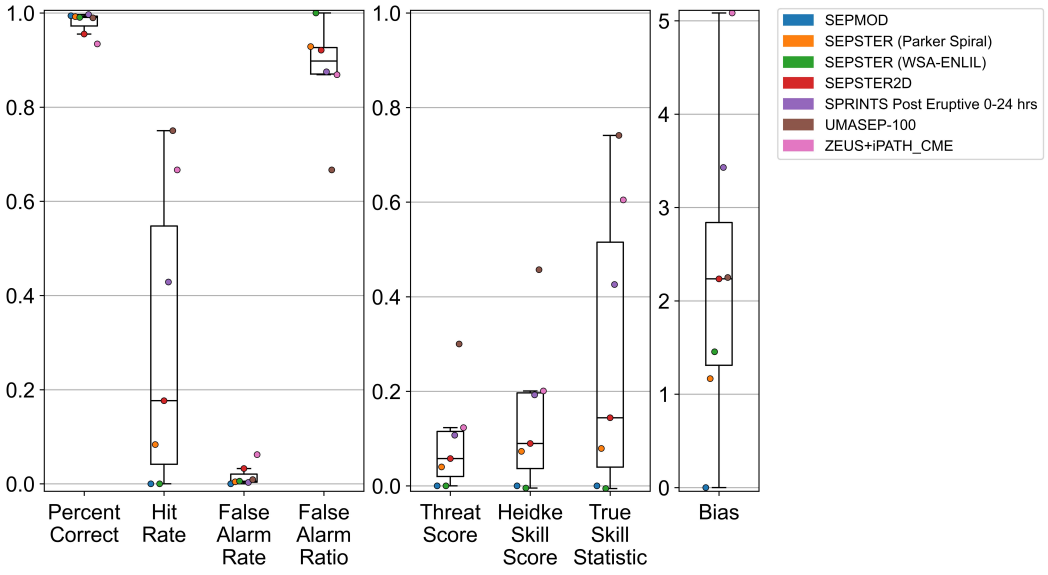


Figure 5.29: All Clear, Post-eruptive, >100 MeV: Summary box plots of SEP Scoreboard All Clear metrics for post-eruptive forecasts associated with >100 MeV integral proton flux with 1 pfu flux threshold.

5.2.2 Probability

Probability forecasts are issued to the Scoreboards by the pre-eruptive models GSU, MAG4, and MagPy and the post-eruptive SPRINTS. Table 5.7 shows the median scores for both model types. SPRINTS is the only post-eruptive probability model evaluated here. It has a low (good) Brier Score, but this is driven by the vast number of correct negatives, however the AUC of 0.70 for >10 MeV and 0.90 for >100 MeV indicates the model does have skill. The pre-eruptive models have a weak performance, although the median AUC of 0.61 and the box plot in Figure 5.31 show small skill above random chance for MagPy and the MAG4 variants.

Scoreboard Median Probability Scores	Scoreboard Post-eruptive >10 MeV	Scoreboard Post-eruptive >100 MeV	Scoreboard Pre-eruptive >10 MeV	Scoreboard Pre-eruptive >100 MeV
Brier Score	0.006	0.002	0.03	-
Brier Skill Score	-	-	-0.04	-
Area Under the Curve	0.70	0.90	0.61	-

Table 5.7: Median probability metrics for the SEP Scoreboards. The Brier Skill Score is calculated using the climatology published in Bain et al. (2021) as a reference, therefore it is not an appropriate climatological reference for models triggered by flares and CME.

In Figure 5.33, the ROC curve for SPRINTS demonstrates skill above the random guess line. Encouragingly, the ROC curves of the pre-eruptive models MAG4 and MagPy, in Figure 5.34, show positive skill over random guess and further, indicate that the All Clear metrics for the LOS versions of MAG4 might increase with a different choice of probability threshold for binary conversion.

In real-time forecasting, it is desirable for issued probabilities to represent the observed frequency of SEPs. The reliability diagrams in Figures 5.35 and 5.36 compared the predicted probabilities with the associated observed SEP frequencies. For >10 MeV, SPRINTS over-predicts the probabilities compared to the observed frequency of SEP events. Figure 5.36 shows that MAG4 variants generally issue forecasts below 50%. The points in the lower right corner are due to a series of 99% forecasts that were issued randomly for a short period of time and are believed to be caused by a temporary bug. The remaining probability forecasts have a tendency to follow the perfect calibration line, indicating that they are fairly calibrated to realistic observed frequencies, however under- and over-prediction is seen for some probabilities.

For the SEP Scoreboards, MAG4 and MagPy demonstrate the best performance with respect to probability. Their issued probabilities are generally aligned with observed SEP frequencies, as demonstrated by their reliability diagrams, and their skill exceeds that of random guess, as shown by the ROC curves and AUC values above 0.5.

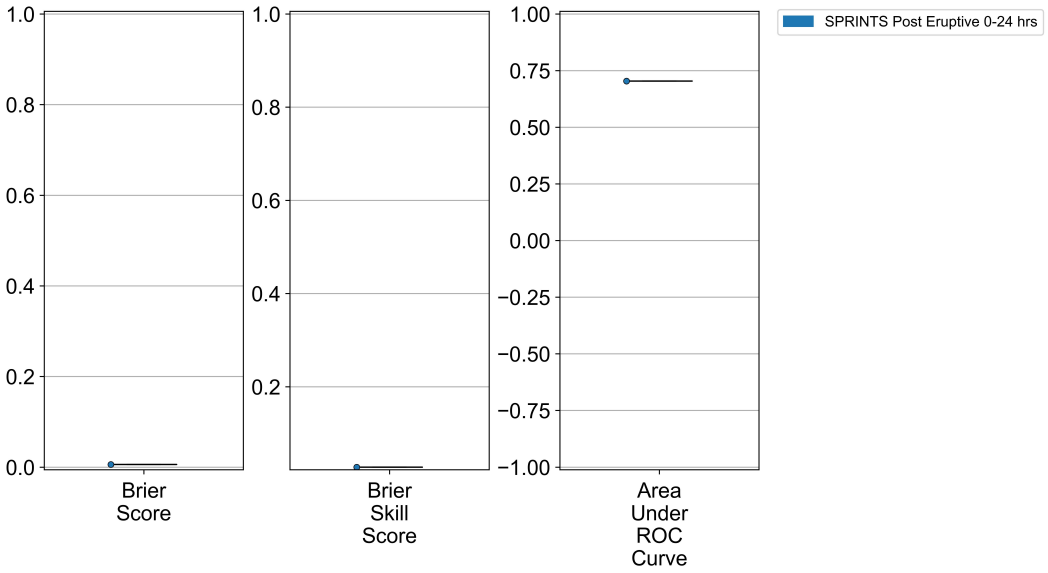


Figure 5.30: Probability, Post-eruptive, >10 MeV: Summary box plots of SEP Scoreboard probability metrics for post-eruptive forecasts associated with >10 MeV integral proton flux with 10 pfu flux threshold.

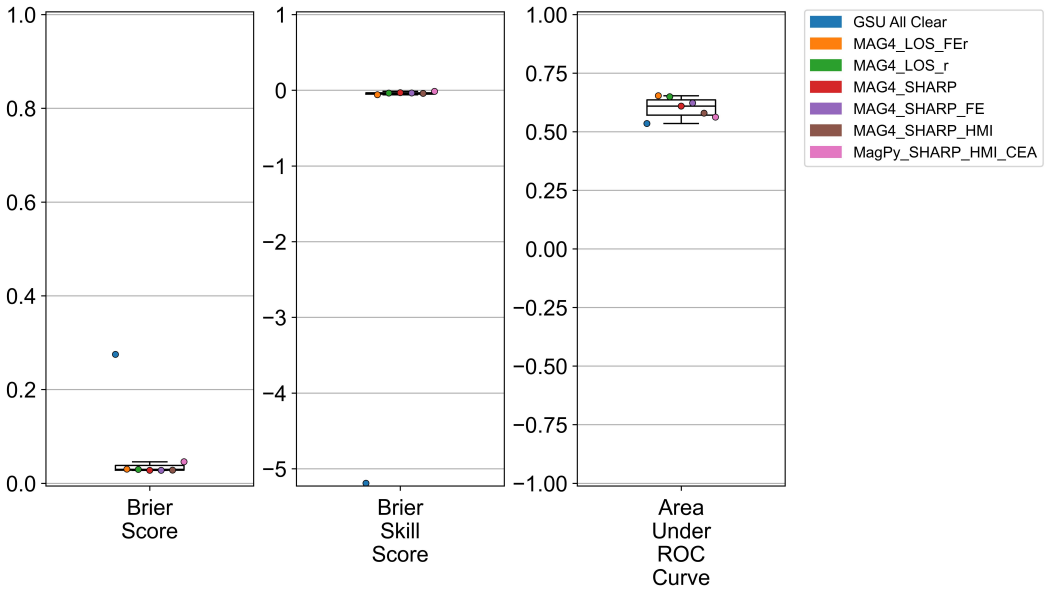


Figure 5.31: Probability, Pre-eruptive, >10 MeV: Summary box plots of SEP Scoreboard probability metrics for pre-eruptive forecasts associated with >10 MeV integral proton flux with 10 pfu flux threshold.

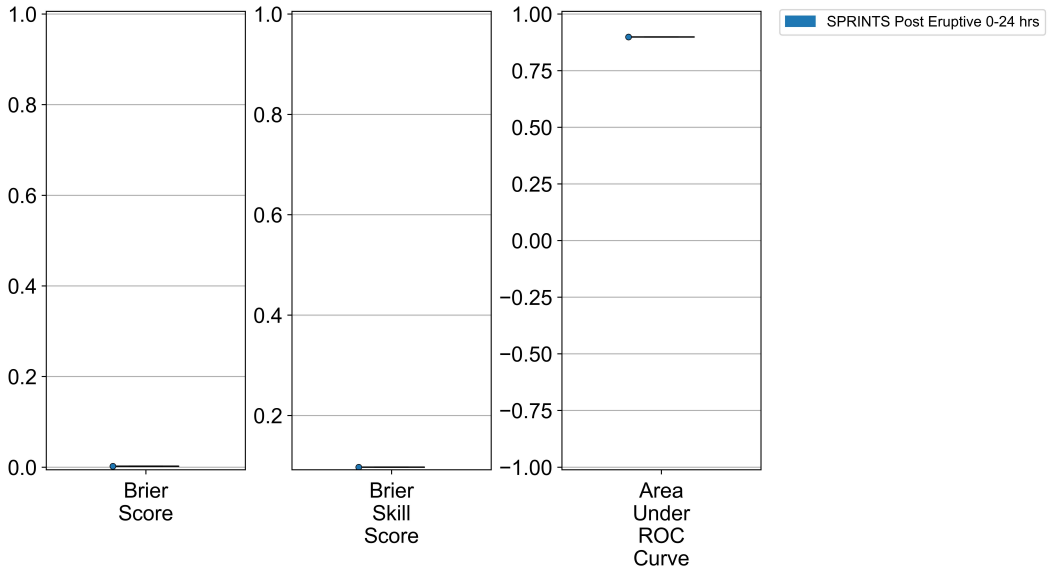


Figure 5.32: Probability, Post-eruptive, >100 MeV: Summary box plots of SEP Scoreboard Probability metrics for forecasts associated with >100 MeV integral proton flux with 1 pfu flux threshold.

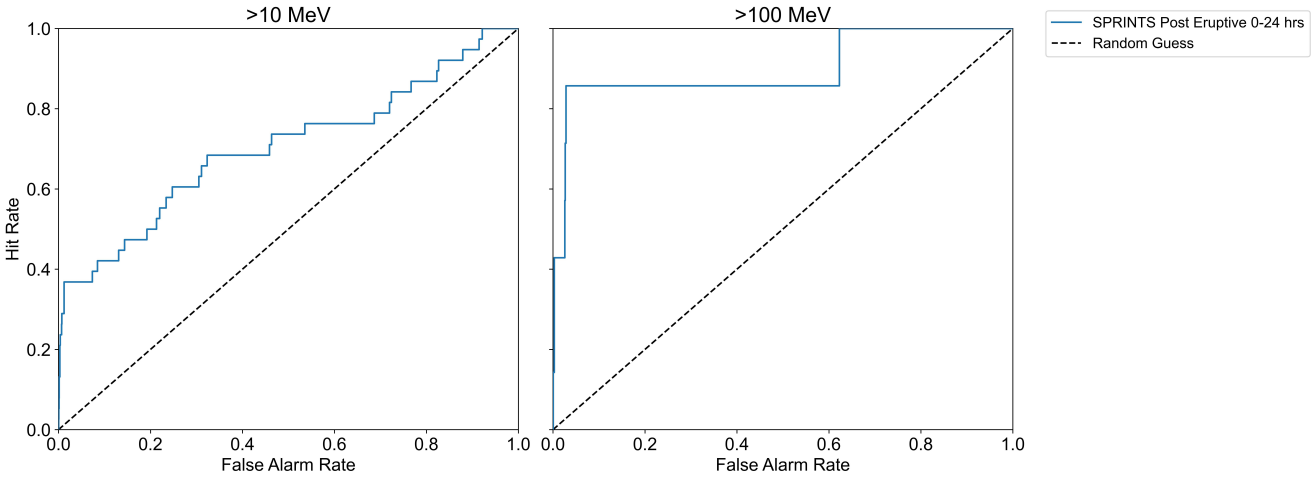


Figure 5.33: Summary ROC curves of SEP Scoreboard SPRINTS forecasts associated with >10 integral proton flux with 10 pfu flux threshold and >100 MeV integral proton flux with 1 pfu flux threshold.

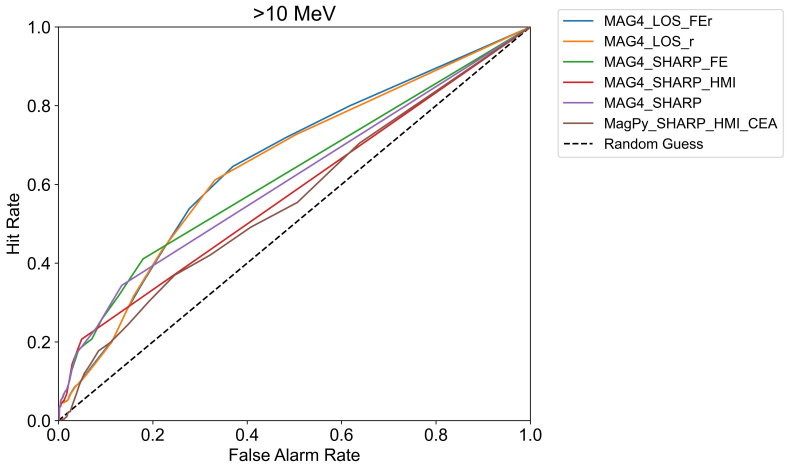


Figure 5.34: Summary ROC curves of SEP Scoreboard MAG4 and MagPy forecasts associated with >10 integral proton flux with 10 pfu flux threshold.

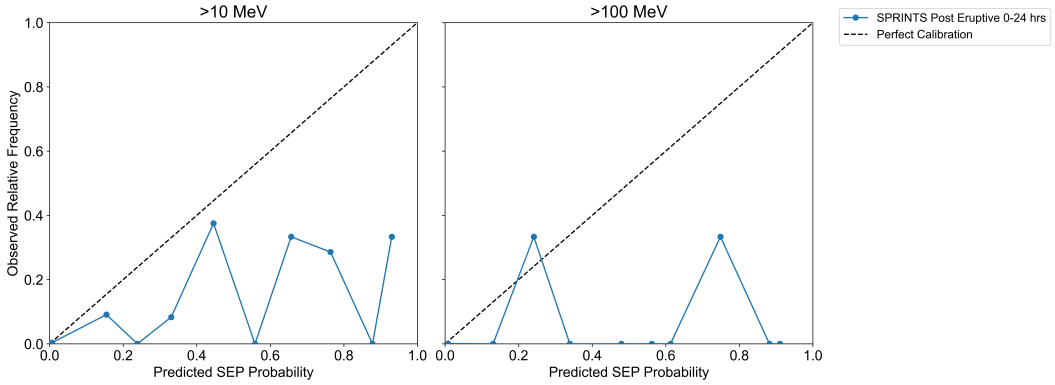


Figure 5.35: Summary reliability diagrams of SEP Scoreboard SPRINTS forecasts associated with >10 integral proton flux with 10 pfu flux threshold and >100 MeV integral proton flux with 1 pfu flux threshold.

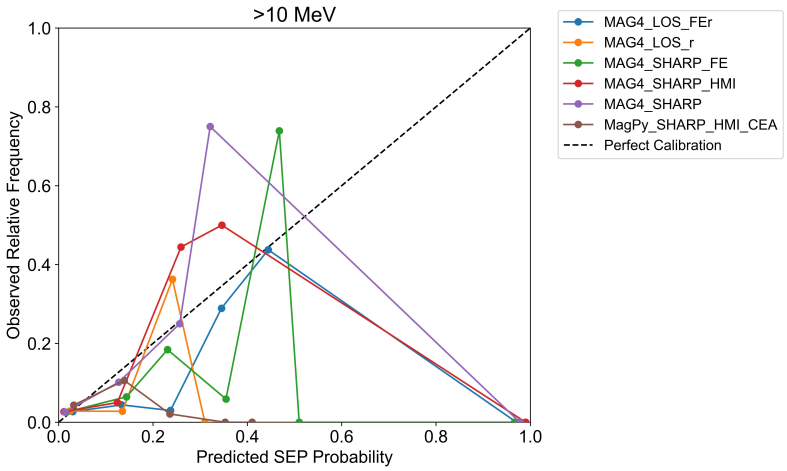


Figure 5.36: Summary reliability diagrams of SEP Scoreboard MAG4 and MagPy forecasts associated with >10 integral proton flux with 10 pfu flux threshold.

5.2.3 Onset Peak and Maximum Flux

Half of the models on the SEP Scoreboards make a peak flux or time profile prediction. SEPSTER was designed to predict the onset peak while SEPSTER2D was developed to predict the maximum flux achieved during an SEP event. SEPMOD and iPATH are physics-based models that track simulated CMEs as they propagate through a solar wind simulation and produce particles with time that travel to Earth. This results in full time profile predictions and both the onset peak and maximum flux are extracted, however only SEPMOD will model ESP enhancements. UMASEP predicts the maximum flux expected to occur in a specific time window that varies according to particle energy and estimated magnetic connectivity. These windows are designed to encompass the onset peak of observed SEP events, but this may not always be the case. All of these models are compared to both the observed onset peak and maximum flux.

Scoreboard Median	Onset Peak		Maximum Flux	
Flux Metrics	>10 MeV	>100 MeV	>10 MeV	>100 MeV
Median Log Error	0.12	−1.12	−0.55	−1.25
Median Absolute Log Error	0.56	1.12	0.68	1.25
Pearson Correlation Coefficient	0.14	0.19	0.22	0.60
Spearman Correlation Coefficient	0.23	0.14	0.21	0.40

Table 5.8: Median metrics for the SEP Scoreboards. The models that contribute to the median scores are listed in Table 5.5. The Pearson Correlation Coefficient is calculated in log space. The Onset Peak refers to the initial rise of the SEP event while the Max Flux refers to the maximum flux value measured during the SEP event.

Table 5.8 reports the median errors for onset peak and maximum flux. For the SEP Scoreboards, the median log errors are within an order of magnitude for both types of peaks and both energies with an underprediction of half an order of magnitude for max flux. The correlation coefficients show that there is little correlation between the predictions and observations, except for >100 MeV maximum flux. It should be kept in mind that a small number (17 or less) of events are included in the correlation values for >100 MeV. Figure 5.37 shows that models generally predict onset peak fluxes in the right order of magnitude, but that the correlation is very low indicating that the event-to-event variability is not being captured. Figure 5.39 shows an overall underestimation of the maximum flux, but this is expected as most of the models are designed to predict the onset peak, with the exception of SEPSTER2D. Figures 5.38 and 5.40 of the log error distributions demonstrate that the models have a spread centered within an order of magnitude of zero. The median absolute log error of 0.56 for onset peak and 0.68 for maximum flux indicate that predictions are generally within an order of magnitude of the observed fluxes. Similarly, log error distributions in Figure 5.38 show that most predictions are within an order of magnitude, but the range of errors extends to 2 or more orders of magnitude. For >100 MeV, Figures 5.41 –5.44 show that the story is the same for

>100 MeV with a tendency to underpredict by more than an order of magnitude. Only iPATH and UMASEP-100 span log errors close to zero. For any individual SEP event, the wide range of possible errors means that peak flux forecasts are unreliable. This is consistent with what was found for SEPVAL.

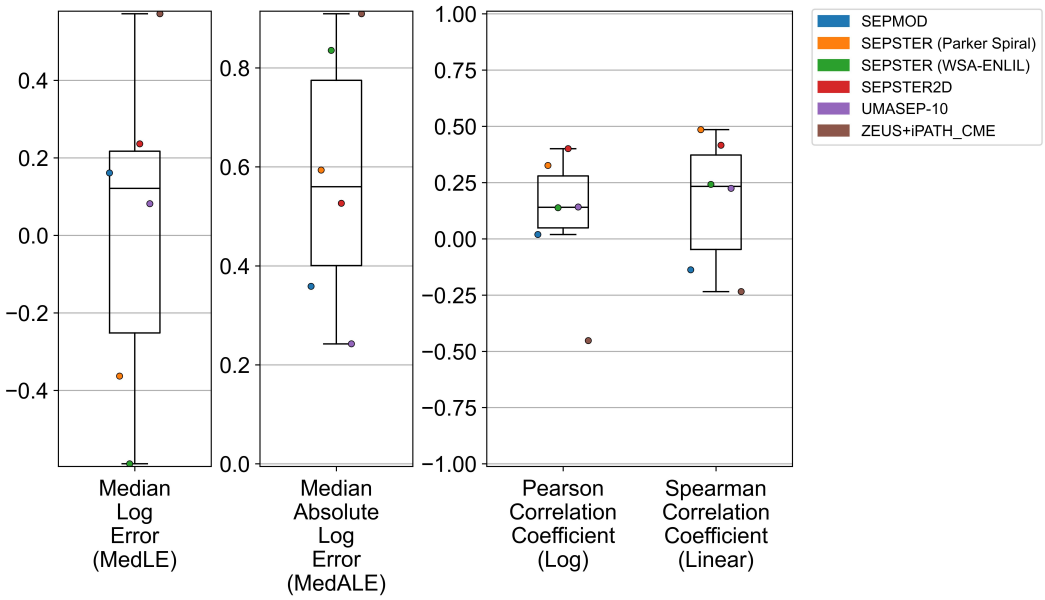


Figure 5.37: Summary box plots of SEP Scoreboard Onset Peak metrics for forecasts associated with >10 MeV integral proton flux with 10 pfu flux threshold.

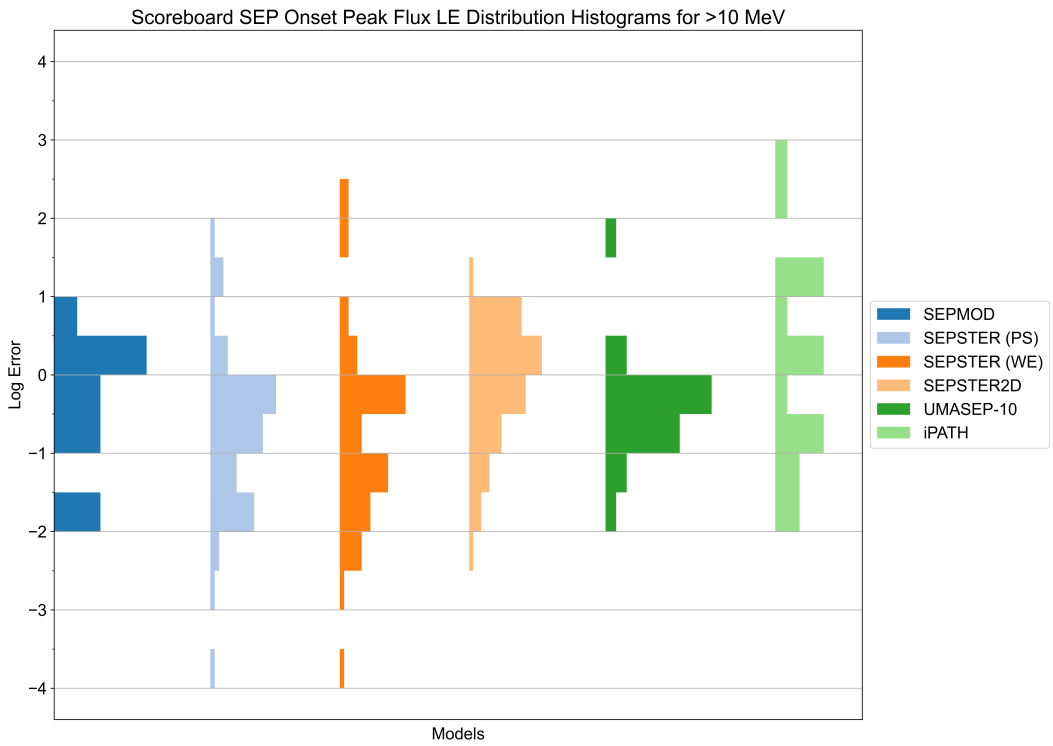


Figure 5.38: Summary histograms of SEP Scoreboard Onset Peak log error for forecasts associated with >10 MeV integral proton flux with 10 pfu flux threshold.

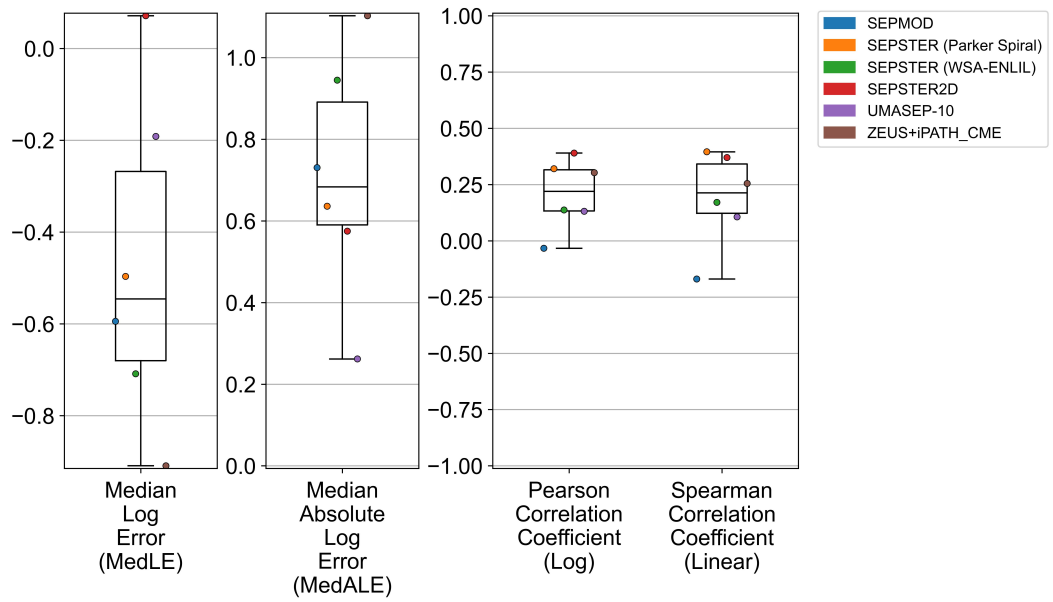


Figure 5.39: Summary box plots of SEP Scoreboard Max Flux metrics for forecasts associated with >10 MeV integral proton flux with 10 pfu flux threshold.

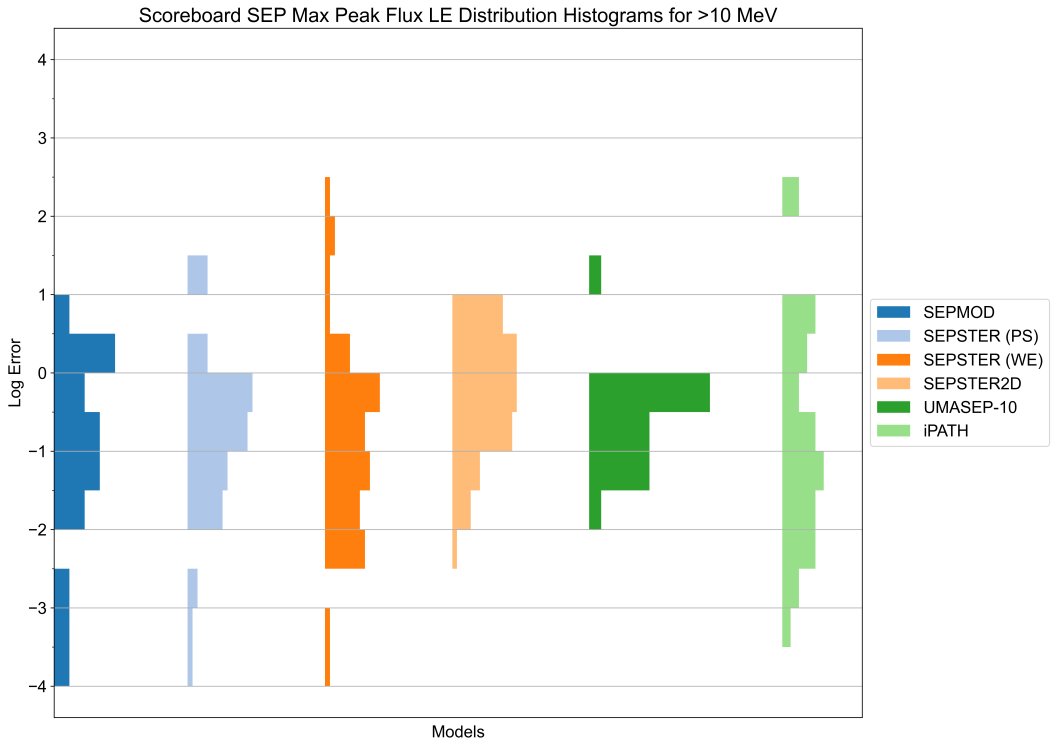


Figure 5.40: Summary histograms of SEP Scoreboard Max Flux log error for forecasts associated with >10 MeV integral proton flux with 10 pfu flux threshold.

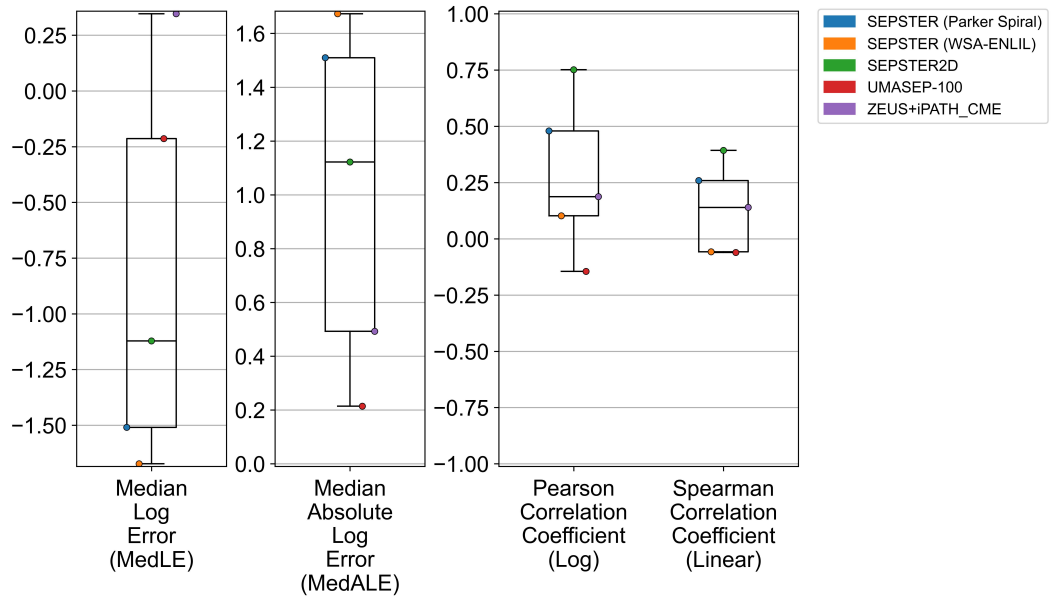


Figure 5.41: Summary box plots of SEP Scoreboard Onset Peak metrics for forecasts associated with >100 MeV integral proton flux with 1 pfu flux threshold.

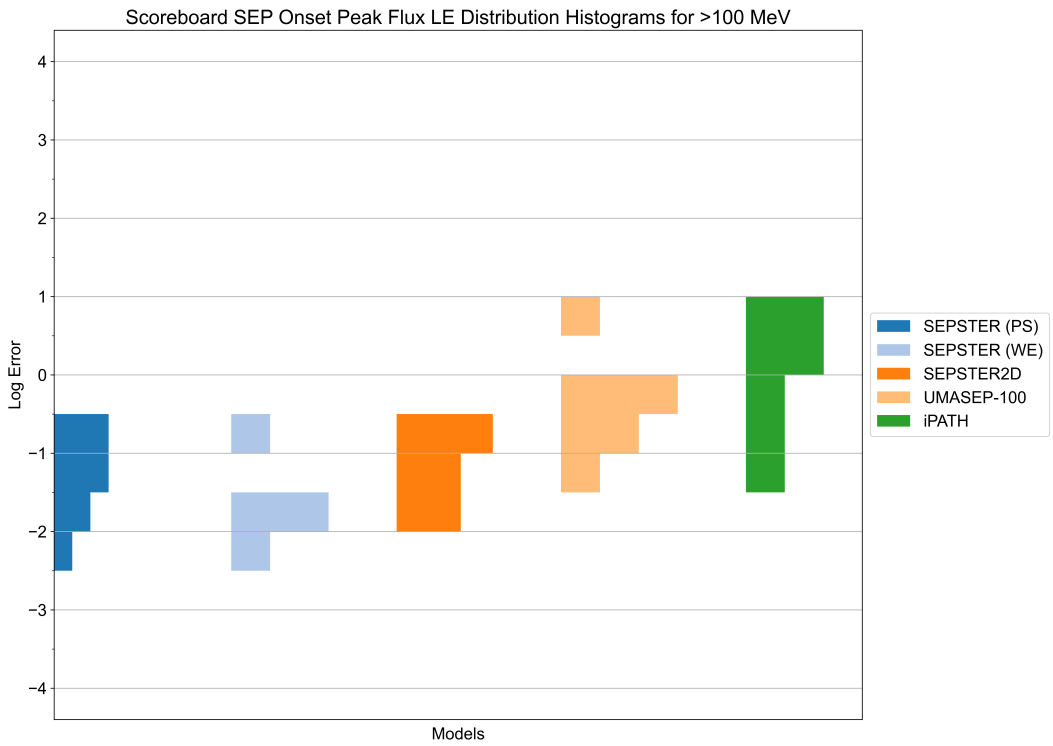


Figure 5.42: Summary histograms of SEP Scoreboard Onset Peak log error for forecasts associated with >100 MeV integral proton flux with 1 pfu flux threshold.

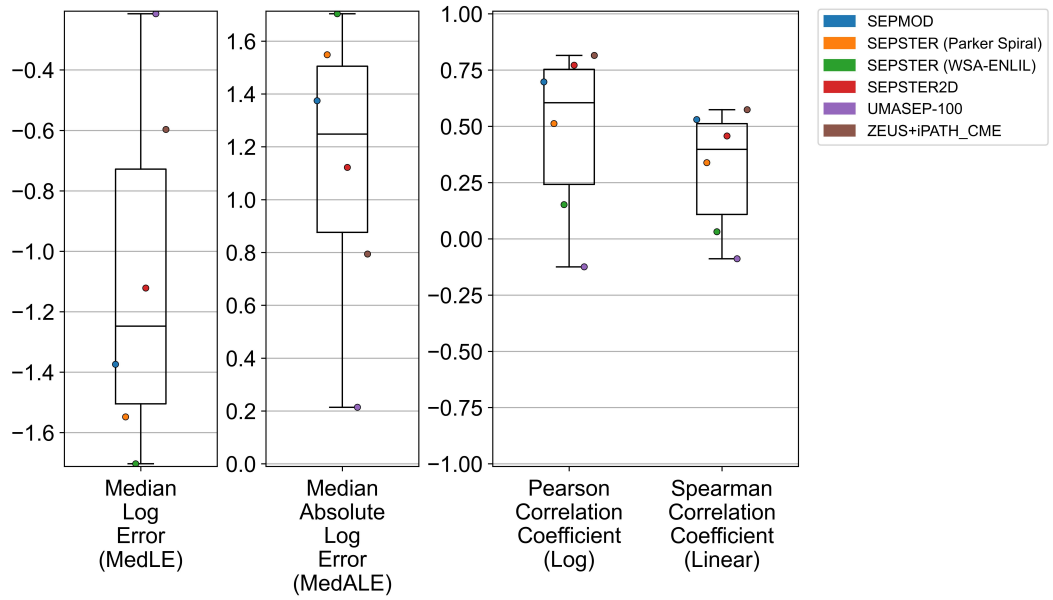


Figure 5.43: Summary box plots of SEP Scoreboard Max Flux metrics for forecasts associated with >100 MeV integral proton flux with 1 pfu flux threshold.

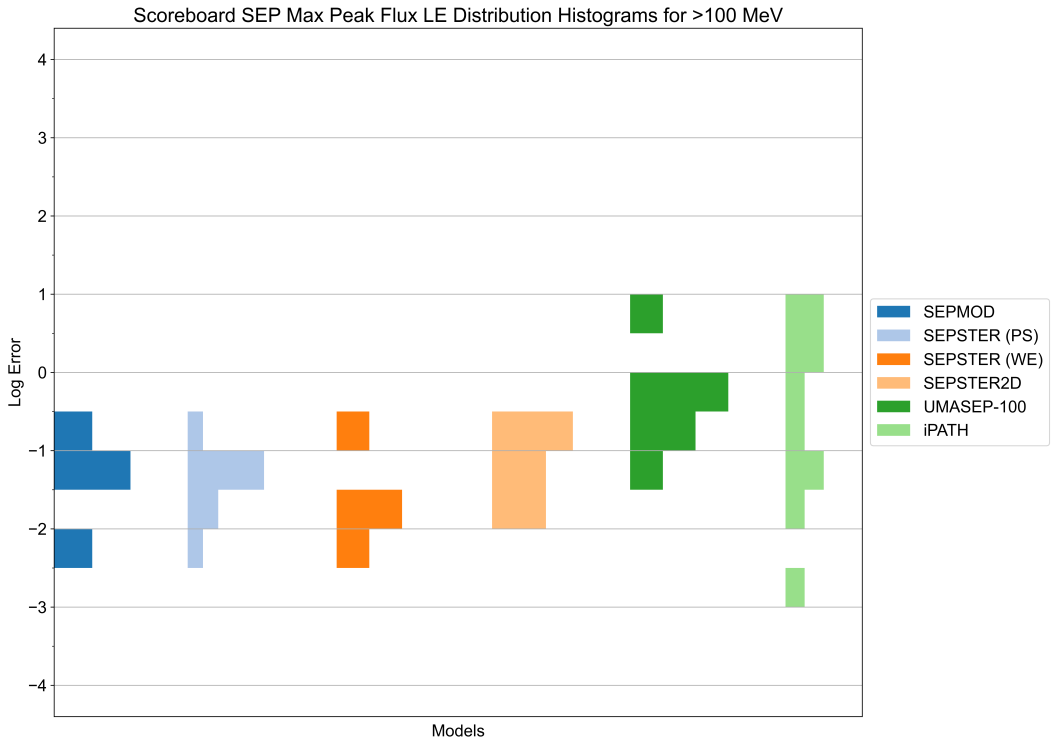


Figure 5.44: Summary histograms of SEP Scoreboard Max Flux log error for forecasts associated with >100 MeV integral proton flux with 1 pfu flux threshold.

5.2.4 Advance Warning Time

The Advance Warning Time (or lead time) evaluates how far in advance a model makes a forecast available to the end-user. It is calculated by comparing the forecast issue time (when the forecast is written to file) and the time that the phenomenon is observed, described further in Section 3.3.4. In this case, we compare the forecast issue time for All Clear forecasts to the observed SEP event threshold crossing time. We also compare the forecast issue time for peak flux forecasts with the observed onset peak time. A positive AWT means that the forecast was issued before the observed start or peak happened and a negative AWT means the forecast was issued after the observed start or peak happened. A negative AWT might occur for models with long run times or due to delayed input data. If the model run time is too long or the input data takes too long to become available, the model may not be able to get the forecast into the hands of the end-user before the observed start or peak has already happened.

Forecasts only provide advance warning if they correctly predict an event will occur, therefore AWT is calculated from the subset of forecasts that were hits. Table 5.9 lists the median AWT per model in hours for >10 MeV events and Table 5.10 lists the median AWT per model in hours for >100 MeV events. The columns labeled "N" indicate the number of hits that were used to calculate the median value.

The labels "Strict" and "First" refer to two different methods implemented in SPHINX to evaluate AWT. For models that produce multiple forecasts ahead of a single SEP event, the "Strict" approach requires that all forecasts leading up to the event are hits. The first forecast in the pre-eruptive sequence of hits is used to calculate AWT. In the "First" approach, the very first hit associated with a SEP is used to calculate AWT, regardless of whether later forecasts may have been misses. The "First" forecast approach is consistent with the method used to calculate deoverlapped All Clear scores, since in that logic, any hit gives the model credit for correctly predicting an event.

In Table 5.9 for >10 MeV, the pre-eruptive models GSU and MAG4 show very long lead times for tens of SEP events. It should be kept in mind that models that produce many false alarms, like GSU and MAG4, may get hits in a random fashion when a series of false alarms happened to "run into" an observed SEP event. AWT can only be appropriately evaluated alongside model skill. Also note that many of the models have a very small number of hits, so the reported AWT reflects only a few events and is not a reliable estimate.

For models with demonstrated skill and a large number of hits, two categories are evident on the SEP Scoreboards – those that require CME parameters and those that do not. SEPSTER, SEPSTER2D, SEPMOD, and iPATH all require CME parameters to be measured by M2M and entered into the DONKI catalog to trigger a run. These models typically issue forecasts an hour or more after an observed SEP event threshold has been crossed. The good news is that these models all predict peak flux and forecasts are typically available prior to the observed onset peak. Figure 5.45 shows that the AWT for CME-input models ranges across positive and negative values, meaning that some forecasts are able to be issued prior to an

observed SEP threshold crossing.

SPRINTS runs automatically after a flare occurs and shows 1.8 hours of warning, however it should be noted that the All Clear and probability sections above show limited skill. UMASEP-10 and HESPERIA-REleASE both ingest *in situ* energetic proton and electron measurements, respectively, to make forecasts. Both models have demonstrated forecasting skill and are often the first models to indicate a change in the space radiation environment on the SEP Scoreboards. UMASEP-10 has a median AWT of 0.74 hours (44 minutes) and HESPERIA-REleASE has median AWT values of 3.4 hours using ACE electron inputs and 1.0 hours using SOHO electron inputs. It should be noted that HESPERIA-REleASE ACE 60-min has a tendency towards false alarms compared to HESPERIA-REleASE SOHO 60-min. The SWPC Day 1 forecasts provide an impressive lead time of over 14 hours, however this forecast does suffer from false alarms. The SWPC Warning product is by far the most reliable as it is typically issued by a forecaster when conditions are obviously leading towards an SEP event. Even so, the median AWT for these warnings is 0.93 hours (56 minutes). It is interesting to note that Figure 5.45 shows a very similar AWT distribution for SWPC Warning and UMASEP-10.

Table 5.10 for >100 MeV tells a similar story – iPATH issues forecasts many hours after a >100 MeV threshold crossing has occurred, and in this case, does not typically provide forecasts before the >100 MeV peak is observed. Only UMASEP-100 and SWPC Warning have enough hits for an informative lead time and both issue forecasts less than 20 minutes prior to the observed event start. Near-relativistic >100 MeV protons can arrive at Earth within 20–30 minutes after an eruption on the Sun, therefore providing advance warning ahead of these energetic events is an extremely challenging task for SEP model forecasting today.

Model	AWT (Strict)	N (Strict)	AWT (First)	N (First)	AWT to Onset Peak (Strict)	N (Strict)	AWT to Onset Peak (First)	N (First)
GSU All Clear	23.1	14	23.3	23	-	-	-	-
MAG4_LOS_FEr	22.5	27	22.6	29	-	-	-	-
MAG4_LOS_r	22.4	23	22.5	28	-	-	-	-
MAG4_SHARP	16.1	10	20.3	23	-	-	-	-
MAG4_SHARP_FE	13.0	15	20.8	24	-	-	-	-
MAG4_SHARP_HMI	16.0	6	19.4	13	-	-	-	-
MagPy_SHARP_HMI_CEA		0	17.6	2	-	-	-	-
SEPMOD	−3.5	8	−3.1	9	−1.5	8	−1.1	9
SEPSTER (Parker Spiral)	−0.61	16	−0.70	18	4.1	16	3.6	18
SEPSTER (WSA-ENLIL)	−0.96	15	−0.96	15	1.4	15	1.4	15
SEPSTER2D	−0.55	27	−0.55	29	3.5	27	3.5	29
SPRINTS Post Eruptive 0-24 hrs	1.81	5	1.8	5	-	-	-	-
SWPC Day 1 (1%)	14.3	28	14.3	28	-	-	-	-
SWPC Day 1 (10%)	14.7	17	14.7	17	-	-	-	-
SWPC Warning	0.93	28	0.93	28	-	-	-	-
UMASEP-10	0.74	23	0.77	24	5.7	23	5.9	24
ZEUS+iPATH_CME	−2.9	11	−2.9	11	2.5	10	2.5	10
HESPERIA REleASE ACE 60-min	-	-	3.4	25	-	-	-	-
HESPERIA REleASE SOHO 60-min	-	-	1.0	25	-	-	-	-

Table 5.9: Median Advance Warning Time in hours for SEP Scoreboard models (>10 MeV). Positive (negative) AWT indicates forecasts were issued before (after) observed threshold crossing or onset peak time. N is the number of forecasts included in the median. Strict specifies that all consecutive forecasts leading up to the threshold crossing or onset peak had to be hits. First specifies the AWT from the first forecast that was a hit. Dash indicates the field is not relevant. Note that AWT is a function of a model's tendency towards hits and false alarms.

Model	AWT (Strict)	N (Strict)	AWT (First)	N (First)	AWT to Onset Peak (Strict)	N (Strict)	AWT to Onset Peak (First)	N (First)
SEPSTER2D	−0.88	1	−4.8	2	3.8	1	−0.36	2
SPRINTS Post Eruptive 0-24 hrs	0.49	3	0.49	3	-	-	-	-
SWPC Warning	0.33	6	0.33	6	-	-	-	-
UMASEP-100	0.29	6	0.29	6	4.4	4	4.1	5
ZEUS+iPATH_CME	−10.0	3	−6.8	4	−2.0	2	−2.1	3

Table 5.10: Median Advance Warning Time in hours for SEP Scoreboard models (>100 MeV). Positive (negative) AWT indicates forecasts were issued before (after) observed threshold crossing or onset peak time. N is the number of forecasts included in the median. Strict specifies that all consecutive forecasts leading up to the threshold crossing or onset peak had to be hits. First specifies the AWT from the first forecast that was a hit. A dash (-) indicates the field is not relevant. Note that AWT is a function of a model's tendency towards hits and false alarms.

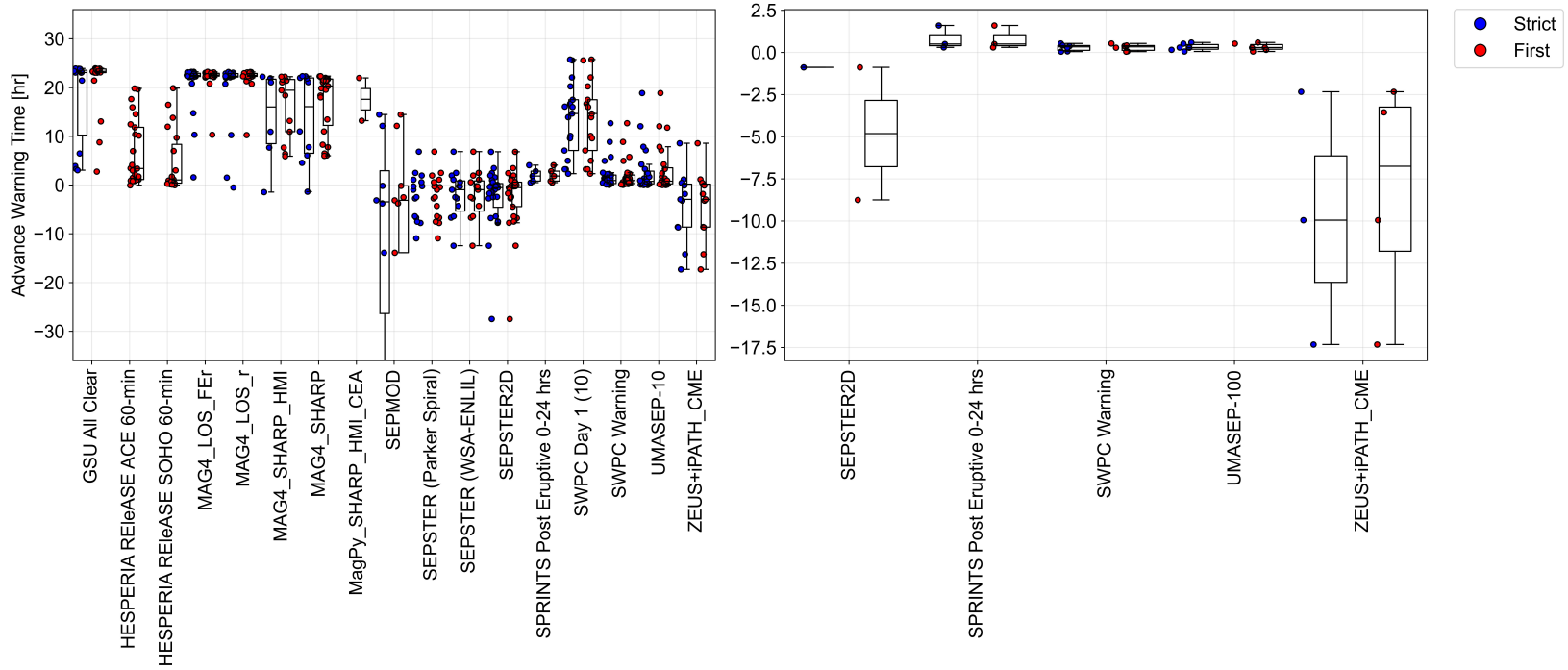


Figure 5.45: Advance Warning Time box plots showing the distribution of AWT per forecast for models on the SEP Scoreboard for >10 MeV proton flux crossing 10 pfu threshold (left) and >100 MeV crossing 1 pfu threshold (right).

5.2.5 SEP Scoreboards Group Results Summary

Models forecasting to the SEP Scoreboards demonstrate the challenge of forecasting in real-time conditions. For >10 MeV All Clear forecasts, the median Hit Rate is only 61%, however the scores vary widely with some models achieving upwards of 70%. The False Alarm Rates are very low, clustered around a median of 4%, however that is not low enough to reduce the number of false alarms to less than the number of observed SEP events. The SEP Scoreboards provide a realistic evaluation of skill for SEP models running in real time. Due to the mediocre hit rate and high numbers of false alarms, HSS is low with a median of 0.28, much lower than is often reported in the literature. The highest skill models for All Clear were UMASEP-10 with a Hit Rate of 69%, False Alarm Rate of 3%, and HSS of 0.46. SEPSTER achieved the second highest HSS of 0.44 with a Hit Rate of 62% and False Alarm Rate of 1.8%. Models must aim to maintain a high Hit Rate while reducing false alarms almost to zero to achieve high skill scores in a realistic climatological scenario for SEP events.

The performance of probability models on the SEP Scoreboards is very low with skill just above random chance. MAG4 and MagPy were able to achieve AUC values greater than random chance. Reliability diagrams demonstrate that these two models are able to produce probability values that trend around the observed frequencies of SEP events.

SEP models have a very difficult time predicting onset peak and maximum flux. For >10 MeV events, SEPMOD, SEPSTER, SEPSTER2D, UMASEP-10, and iPATH, have biases for log error close to zero. The distribution of errors extends across multiple orders of magnitude for most models. For >100 MeV, the overall story is the same with an increased bias towards underprediction. SEP Scoreboard forecasts are able to trend around observed onset peak and maximum flux values in a statistical sense, but for any given forecast, the error may be up to 3 orders of magnitude, making it very difficult to use peak flux forecasts in operations.

The SEP Scoreboards provide the opportunity to quantify Advance Warning Time, the lead time provided by the model. The AWT reported here is the difference between the forecast issue time (when the forecast is written to file) and the observed SEP event threshold crossing time. Additionally, AWT to onset peak time is calculated for models that give peak flux forecasts. Advance warning must be interpreted in the context of model skill and the likelihood for hits and false alarms. Models that require CME parameters as input typically have negative AWT due to the current delay in receiving coronagraph imagery as well as operational constraints in measuring CMEs in a timely manner (i.e. M2M not 24/7), although some forecasts are able to be issued ahead of SEP start times. UMASEP and HESPERIA-REleASE provide about an hour advance warning, which is similar to the SWPC Warnings issued when an SEP event is imminent.

6 Individual Model Validation Results

6.1 SWPC as a baseline model

NOAA SWPC is the official space weather forecasting service for the United States. SWPC forecasters use a wide variety of operational assets to monitor the Sun 24/7, and produce dozens of forecasting products at a regular cadence touching upon all aspects of space weather, including solar active regions, solar flares, radio bursts, geomagnetic storms, and SEPs. SWPC has a long history of collaboration with SRAG dating back to the Apollo era. As such, their forecasting skill is a natural reference point from which to compare computer models with similar objectives.

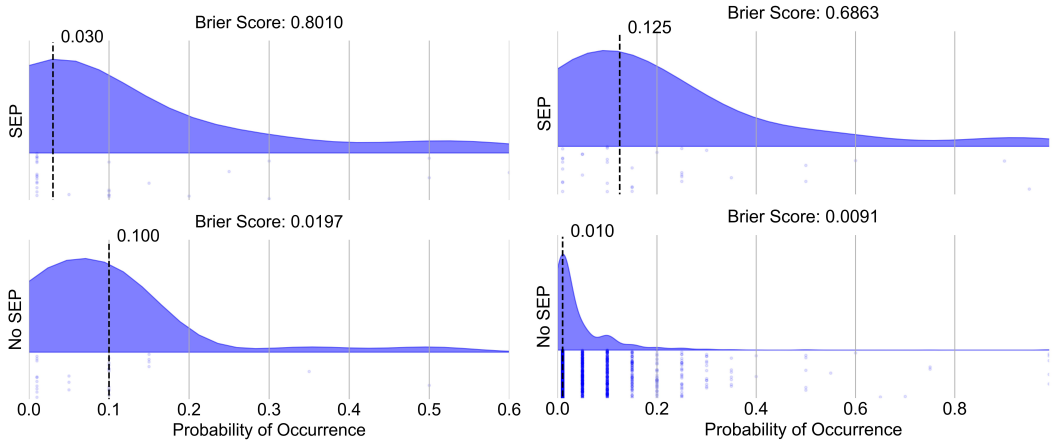
SWPC forecasting products range from free-form discussions of overall space weather conditions, to quantitative forecasts of specific events, with rigorous definitions and meaning. Two of these forecasting products are evaluated here using the SPHINX framework in order to serve as a basis of comparison for the largely automated computer models evaluated throughout this report. SWPC Day-1 forecasts, discussed in Section 6.1.1, are daily probabilistic proton event forecasts for SPEs, or "S1 Events" using the SWPC solar radiation storm scale. SWPC Warnings, discussed in Section 6.1.2, are issued when an SPE or ESPE appears to be imminent. Both of these types of forecasts are extensively validated in the analysis of Bain et al. (2021). However, our analysis is necessary in order to (1) evaluate SWPC forecasts for the time frame of Scoreboard operation (2019–2024; the ascending phase of solar cycle 25) and for the specific set of SEPVAL event periods, and (2) to use exactly the same SPHINX forecast matching logic that will be applied to other models in this analysis (see Section 3.2). This provides us with a fair set of performance metrics to use as a basic of comparison for the Scoreboard models. However, those differences will lead to discrepancy with respect to results published in Bain et al. (2021), which covered the entire solar cycles 23 and 24. In particular, readers are cautioned that our "All Clear" analysis in the following section uses a different probability threshold than what was used in Bain et al. (2021), resulting in skill scores that are not comparable.

6.1.1 SWPC Day-1

SWPC Day 1 forecasts were gathered from archives of the Report of Solar and Geophysical Activity (RSGA). This product is issued daily at 22:00 UTC and contains a probability forecast for SPEs to occur on days 1, 2, and 3 following midnight (00:00 UTC) after the forecast was issued. We only consider the forecast for the following day, Day-1. SWPC never fails to issue a forecast, however our collection of archival data was missing forecasts for 18 days, resulting in the difference between N Days and N forecasts in Table 6.1. Furthermore, SPHINX matching logic rejects forecasts that are issued during an event, and this occurred for 38 forecasts, resulting in a total of 2135 forecasts considered. This is a crucial difference between our analysis and that of Bain et al. (2021), where SWPC often issues forecasts of 99% probability of an event to occur the following day when an event is ongoing at the issue time of 22:00 UTC. These persistence forecast "hits" are not operationally useful for SRAG

SWPC Day-1	
Characteristic	> 10 MeV
First Forecast	2019-01-02
Last Forecast	2024-12-31
N Days	2191
N Forecast Days	2172
N SEP Days	34
Forecast Cadence	daily
Prediction Window	24 hr
N forecasts	2173
N matched w/events	34
Imbalance (raw)	62.9
Imbalance (days)	62.9

Table 6.1: SWPC Day-1 validation characteristics.



(a) SWPC Day-1 SEPVAL

(b) SWPC Day-1 Scoreboard

Figure 6.1: Forecasted probabilities and Gaussian kernel density estimates for SWPC Day-1 from SEPVAL and the SEP Scoreboard, separated by event periods and non-event periods. Considers >10 MeV integral proton flux with a 10 pfu flux threshold.

and are not considered by SPHINX.

Figure 6.1 shows the distributions of probability forecasts issued by SWPC Day-1. We see good performance in distinguishing SEP event periods from non-event periods in real-time operations during the Scoreboard period, with a significantly higher median probability of 12.5% for event periods over the 1.0% median during non-event periods. The distribution for the non-event periods is also very narrow, with the vast majority of forecasts being either 1% or 5%. Note that SWPC Day-1 forecasts are typically quantized at 5% intervals, with the exception on their preference to issue a 1% forecast instead of a 0% forecast. By comparison, the event

period distribution is much broader and has an extended tail of higher probability forecasts. Note that the fatter tail is only relative to the peak: there are many more probability forecasts issued during non-event periods, as seen in the scatter plot below the distribution. But relative to the overwhelming peak of 1% forecasts the high-probability tail is much smaller.

The performance for the SEPVAL dataset was not as good. Figure 6.1a shows that *higher* probability forecasts were issued during non-event periods than event periods. This appears to be due to a relatively larger number of 10% forecasts for this particular challenge set. This outcome would seem to indicate difficulty in distinguishing SEP-producing solar conditions when many of the underlying characteristics of the event and non-event periods are the same. Recall that the SEPVAL non-event periods were intentionally biased to contain trigger events (flares and CMEs) that have characteristics similar to those in the event period class (see Section 4). It is likely that this bias also resulted in similarities in other characteristics that SWPC considers when making the Day-1 forecast, such as AR sizes and complexity, recent flaring rates, etc. The Scoreboard results can be interpreted as evidence of skill in distinguishing active, SEP-producing periods from much more frequent quiet periods, while the SEPVAL results indicate difficulty distinguishing SEP-producing active periods from non-SEP producing active periods.

SWPC Day-1 Probability	SEPVAL > 10 MeV ($N = 58$)	Scoreboard > 10 MeV ($N = 2135$)
Brier Score	0.41	0.02
Brier Skill Score	0.12	−0.25
Brier (SEP)	0.801	0.686
Median P SEP	0.030	0.125
Area Under the Curve	0.48	0.78

Table 6.2: SWPC Day-1 probability metrics.

Table 6.2 summarizes probability performance metrics for SWPC Day-1. The performance follows from the quality of the probability distributions described above: performance of the continuous set of forecasts issued during the Scoreboard time period is better than for the limited set of 58 event and non-event periods for SEPVAL. The Brier Score for the Scoreboard is close to the ideal value of 0, dominated by the correct issuance of low probability forecasts during the vast majority of quiet days. The Brier Skill Score for the Scoreboard is negative, indicating *worse* Brier Score performance than a consistent climatological forecast of 3.3%, the value found by Bain et al. (2021) for solar cycle 24. As described above, the median probability issued prior to Scoreboard events was 12.5% and well separated from the median 1% issued for non-event periods. The Area Under the Curve was reasonably high, 0.78.

Figure 6.2 shows the reliability diagram for SWPC Day-1. The log-scale histogram included in the plot shows a nearly power-law decrease in the numbers of higher probability forecasts up to 40%, after which even higher probability forecasts

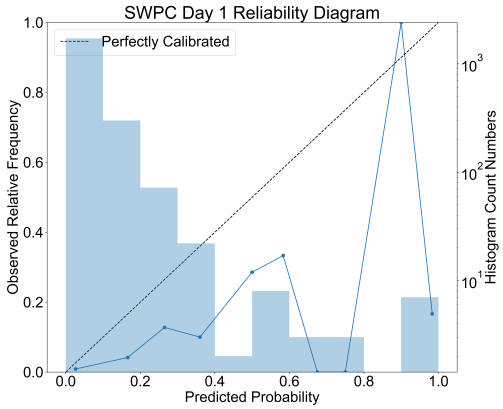
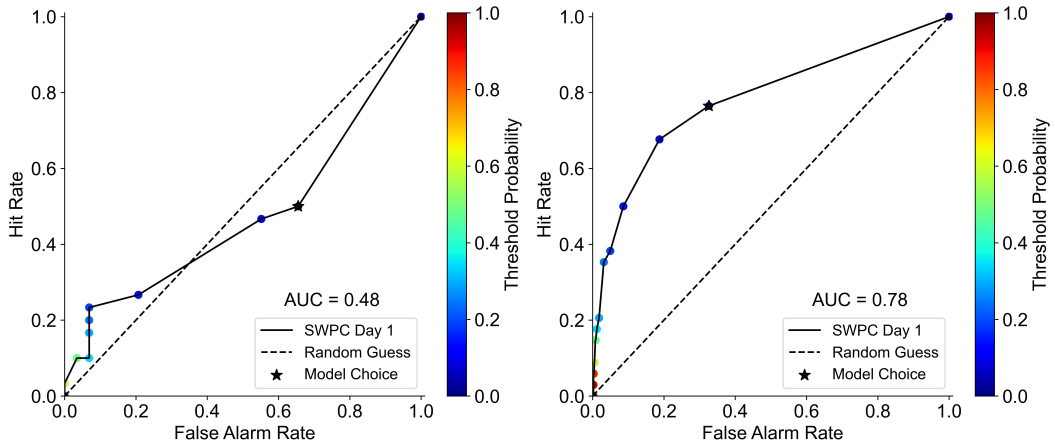


Figure 6.2: Reliability diagram for SWPC Day-1 from the SEP Scoreboard. Considers >10 MeV integral proton flux with a 10 pfu flux threshold.

are very infrequent. The issued probabilities are higher than the observed frequency of events, indicating that SWPC systematically overestimates the likelihood of an event. This result was also seen in Bain et al. (2021) for lower probability forecasts, but there the situation improved for high probability forecasts $P > 80\%$. We do not see a similar improvement due to our removal of persistence forecasts issued during an event.



(a) SWPC Day-1 SEPVAL

(b) SWPC Day-1 Scoreboard

Figure 6.3: ROC curves for SWPC Day-1 from SEPVAL and the SEP Scoreboard. Considers >10 MeV integral proton flux with a 10 pfu flux threshold. Note that the "Model Choice" probability threshold was arbitrarily set at 1%. SWPC does not issue All Clear forecasts.

Figure 6.3 shows the ROC curves for SWPC Day-1. Following the trends in behavior observed from inspecting the event/non-event probability distributions in

Figure 6.1, we see that the ROC curve is quite reasonable for the Scoreboard, but very poor for SEPVAL. The Scoreboard curve, covering the rise phase of solar cycle 25 (2019–2024) is not as good as the aggregate curves presented in Bain et al. (2021) for solar cycles 23 & 24, surely due to our rejection of forecasts issued during ongoing SEPs. That work obtained AUC scores of 0.88 and 0.90 respectively[15], compared to our result of 0.78. Nonetheless, our results show that a relatively good "All Clear" model can be derived from SWPC Day-1 probability forecasts.

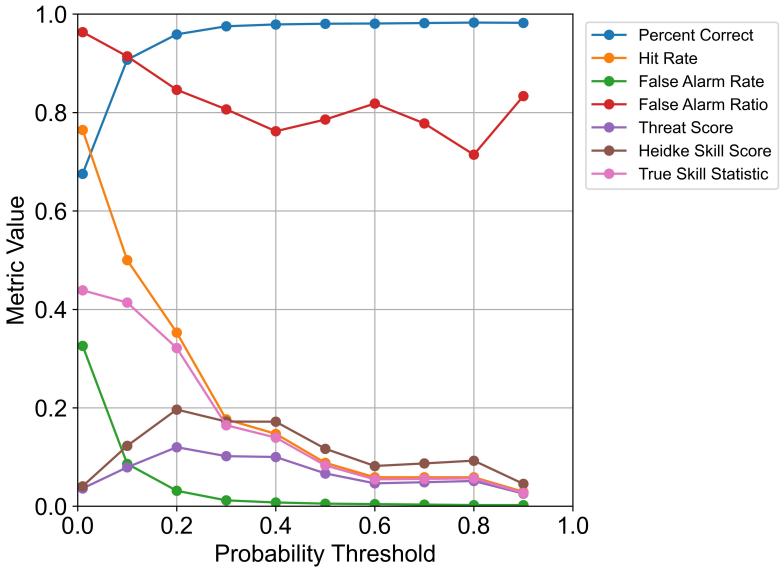


Figure 6.4: SWPC Day-1 All Clear metrics as a function of the chosen probability threshold.

SWPC does not issue an All Clear forecast product. Therefore, for the purposes of comparison to other All Clear forecasts in this report, we construct a SWPC All Clear model by choosing a threshold on SWPC Day-1 probabilities. This exercise is a case study in the tradeoffs that must be made and kept in mind when comparing models on the basis of metrics of binary forecasts like All Clear. Figure 6.4 shows how our chosen set of metrics change with the choice of probability threshold. Metrics values are tabulated up to 50% in Table 6.3, with the best metric values in bold. As can be seen from the table, there is no choice of threshold that optimizes all metrics simultaneously — significant tradeoffs are involved that ultimately depend on whether the application prefers to minimize misses ($1 - \text{Hit Rate}$) or false alarms. For the lower range of probability thresholds, the Hit Rate drops dramatically, as does the False Alarm Rate. Note how quickly the True Skill Statistic becomes equivalent to the Hit Rate as the False Alarm Rate trends to zero, as can be understood from its definition given in Equation 3.8. For the purposes of model comparison in this work we choose a middle point in the metric trends at a probability threshold

[15]Note that Bain et al. (2021) published the ROC skill score, ROCSS = 2AUC − 1. We have converted *ROCSS* back to *AUC* for comparison.

	This Work SC25 (2019–2024)						Bain 2021 SC23	Bain 2021 SC24
Threshold	1%	**10%**	20%	30%	40%	50%	50%	50%
Percent Correct	0.68	0.91	0.96	0.98	0.98	**0.98**	0.97	0.99
Hit Rate	**0.76**	0.50	0.35	0.18	0.15	0.09	0.47	0.62
False Alarm Rate	0.33	0.09	0.03	0.012	0.008	**0.005**	0.002	0.004
False Alarm Ratio	0.96	0.91	0.85	0.81	**0.76**	0.79	0.07	0.16
Threat Score	0.04	0.08	**0.12**	0.10	0.07	0.05	0.45	0.55
Bias	21.0	5.82	2.29	**0.91**	0.62	0.41	0.50	0.73
HSS	0.04	0.12	**0.20**	0.17	0.17	0.12	0.61	0.70
TSS	**0.44**	0.41	0.32	0.16	0.14	0.08	0.47	0.61

Table 6.3: SWPC Day-1 All Clear metrics from Scoreboard forecasts by selection of probability threshold. The best value for each metric is bold, as is our choice for probability threshold. Results are compared to those of Bain et al. (2021) where hits attributed to persistence forecasts issued during an ongoing event result in a large difference in the metrics compared to our work which omits those forecasts.

.

of 10%. At this value, the Hit Rate (50%) is near the middle of its range, the False Alarm Ratio remains high at 91%, but it is not possible to drive it too low in any case, and the TSS (0.41) is near its high end while the HSS (0.12) is near the middle of its range.

As discussed in the introduction, the differences in our dataset and approach make our metrics uncomparable to those published in Bain et al. (2021). In that work, a threshold analysis analogous to what is shown in Figure 6.4 showed that the Hit Rate declined slowly from $\sim 70\%$ to $\sim 50\%$ for probability thresholds ranging from 20% to 90%. HSS remains at relatively high values >0.60 throughout that range, and False Alarm Ratios fall below 20%. Table 6.3 compares the metrics directly. Our results indicate that the combination of high Hit Rate, low False Alarm Ratio, and high HSS are unattainable when omitting the persistence forecasts issued during an event.

> 10 MeV

SEPVAL

	Observed Yes	Observed No	Sum
Pred. Yes	8	6	14
Pred. No	22	23	45
Sum	30	29	59

Scoreboard

	Observed Yes	Observed No	Sum
Pred. Yes	17	181	198
Pred. No	17	1920	1937
Sum	34	2101	2135

Table 6.4: Contingency tables for SWPC Day-1 at a 10% threshold for SEPVAL and the SEP Scoreboard.

Table 6.4 gives the contingency table for SWPC Day-1 forecasts at a >10% probability threshold for both the SEPVAL and Scoreboard datasets, while Table 6.5 gives the resulting metrics. For the Scoreboard, as discussed above, we chose

SWPC Day-1 (10%) All Clear	SEPVAL > 10 MeV ($N = 59$)	Scoreboard > 10 MeV ($N = 2135$)
Percent Correct	0.52	0.91
Hit Rate	0.27	0.50
False Alarm Rate	0.21	0.09
False Alarm Ratio	0.43	0.91
Bias	0.47	5.82
Threat Score	0.22	0.08
HSS	0.06	0.12
TSS	0.06	0.41

Table 6.5: SWPC Day-1 (10%) All Clear metrics.

a threshold that provides a provide medium Hit Rate (50%), which results in a relatively high False Alarm Ratio and low HSS. For the SEPVAL dataset the indicated performance is worse, with a Hit Rate of only 27% and lower skill scores of HSS and TSS at 0.06. The significant decline in skill scores for the SEPVAL challenge set is surprising given the apparent quality of the SWPC probability forecasts discussed above. One possible explanation is the relatively active non-event period selections for SEPVAL are particularly challenging for SWPC to distinguish, while SWPC excels at distinguishing quiet periods from active periods.

6.1.2 SWPC Warnings

SWPC Warnings are issued when a >10 MeV GOES proton flux is expected to exceed 10 pfu. SWPC only issues warnings when a threshold crossing is expected and never when conditions are expected to remain quiet. The contingency table therefore has only hits and false alarms, as this type of forecast will not result in correct negatives. If a warning was issued after an event began, SPHINX does not evaluate it. This is why the contingency table contains less than the 37 SEP events that were observed during the evaluated period. This is different than the approach of Bain et al. (2021), who counted warnings issued after the threshold crossing as a "miss". SWPC sometimes issues multiple warnings for the same SEP event. We have separately "deoverlapped" these forecasts by removing the duplicity.

All Warnings

	Observed		
	Yes	No	Sum
Pred. Yes	38	14	52
Pred. No	–	–	–
Sum	38	14	52

Deoverlapped

	Observed		
	Yes	No	Sum
Pred. Yes	29	12	41
Pred. No	–	–	–
Sum	29	12	41

Table 6.6: Contingency tables for SWPC Warnings on the SEP Scoreboards for >10 MeV exceeds 10 pfu. All warnings (left). Deoverlapped warnings (right).

	Observed		
	Yes	No	Sum
Pred. Yes	28	1	29
Pred. No	5	29	34
Sum	33	30	63

Table 6.7: Contingency table for SWPC Warnings for the SEPVAL events. Here the choice not to issue a warning was interpreted as a "No Event" forecast.

Table 6.6 reports the contingency tables for all of the SWPC Warnings issued to the SEP Scoreboards (left) and the result from removing duplicate (right). Table 6.7 gives the contingency table for SWPC warnings issued for 25 warnings issued during the SEPVAL events. For the purposes of a more complete comparison with flare/CME-triggered post-eruptive models that participate in SEPVAL we treat SWPC Warnings differently here: failure to issue a warning is counted as an "All Clear" forecast. This is analogous to assuming that whatever particle flux increases, CMEs, or flares that were observed by SWPC did not warrant issuing a warning and a threshold crossing was not expected to happen.

Table 6.8 gives the All Clear metrics for SWPC Warnings for SEPVAL and the Scoreboard. For the Scoreboard, only the percent correct and the False Alarm Ratio can be computed, as there is no concept of an "All Clear" warning. The metrics simply say that 71% the SWPC warnings were followed by events. These

SWPC Warning	This Work		Bain 2021	
Dataset	SEPVAL	Scoreboard	SC23	SC24
N	63	41	83	54
Percent Correct	0.90	0.71	0.77	0.76
Hit Rate	0.85	-	-	-
False Alarm Rate	0.03	-	-	-
False Alarm Ratio	0.03	0.29	0.23	0.24
Bias	0.88	-	-	-
Threat Score	0.82	-	-	-
HSS	0.81	-	-	-
TSS	0.82	-	-	-

Table 6.8: SWPC Warning All Clear metrics for SEPVAL and the SEP Scoreboard, as compared to results from Bain et al. (2021) adapted to our Scoreboard methodology, see text.

results can be compared to those from Bain et al. (2021), where we have ignored their misses that were counted for Warnings issued after threshold crossing since SPHINX ignores such forecasts. For SEPVAL we have a complete set of metrics, and the performance is excellent: 90% of the challenge periods were forecast correctly, with skill scores > 0.80. The False Alarm Ratio was an exceptionally low 3%. SWPC is highly accurate in issuing Warnings prior to large events represented in the SEPVAL challenge.

Advance warning time to event threshold crossing was also analyzed for the Scoreboard events, producing the following results:

- Mean AWT = 1.9 hours (29 events)
- Median AWT = 0.93 hours (29 events)
- Shortest AWT = 0.033 hours (2 minutes)
- Longest AWT = 12.7 hours (12 hours 42 minutes)

Warnings were issued after the threshold was already crossed for 9 events. These would have negative AWT, but are not included in the metrics above. These warning times are comparable to Bain et al. (2021), who found median AWT of 57 min for SC23 and 88 min for SC24.

6.2 MAG4

MAG4	LOS_FEr	LOS_r	SHARP_FE	SHARP	SHARP_HMI
First Forecast	2021-02-09	2021-01-01	2021-04-28	2021-04-28	2021-04-28
Last Forecast	2024-11-27	2024-11-27	2024-11-27	2024-11-27	2024-11-27
N Days	1388	1426	1310	1310	1310
N Forecast Days	1276	1281	1239	1239	1239
N SEP Days	35	35	34	34	34
Forecast Cadence	1 hour	1 hour	1 hour	1 hour	1 hour
Prediction Window	24 hours	24 hours	24 hours	24 hours	24 hours
N forecasts	28027	28040	19214	19531	19813
N matched w/events	793	793	511	521	532
Imbalance (raw)	34.3	34.4	35.4	36.5	36.2
Imbalance (days)	35.5	35.6	36.6	35.4	35.4

Table 6.9: MAG4 Scoreboard validation characteristics.

The Magnetogram Forecast (MAG4) model produces a probabilistic forecast for several categories of solar events based on magnetogram inputs. These categories include M+X-class flares, CMEs, fast CMEs, and SPEs. Our validation is only concerned with its SPE forecasts. A previously published validation effort in Falconer et al. (2014) and summarized in Whitman et al. (2023) evaluates flare forecasts.

MAG4 was developed by David Falconer at the University of Alabama in Huntsville and has been described in several publications (see Falconer et al., 2011, 2012, 2014, and references therein). There has been a long history of collaboration with SRAG in the development of this model, starting in 2011 and extending into the ongoing development of MagPy described in the following section. First deployed on the SEP Scoreboard in 2020, there is a large volume of real-time forecasts available for validation analysis (over 114,000 across five variants, see Table 6.9). Early MAG4 forecasts did not conform to the SPHINX file formatting standards and were rejected, with the first valid forecast appearing in Feb 2021.

MAG4	LOS_r	SHARP_HMI
Event Periods	33	33
Non-Event Periods	23	24
Total	56	57
Imbalance	0.70	0.73

Table 6.10: MAG4 SEPVAL validation characteristics.

Support for this model has waned in recent years as focus has shifted to MagPy. As a result, there was no active model developer participating in the SEPVAL challenge. The forecasts for the SEPVAL challenge periods were gathered from a previous large-scale production of forecasts from archival magnetogram data, covering the period 09/2012 to 03/2023, in addition to many of the periods being run on internal CCMC servers. MAG4 forecasts for this period are available on the ISWA data tree, and from this source forecasts covering a subset of the SEPVAL challenge periods were gathered. Table 6.10 shows the number event and non-event periods

that were obtained. The imbalance being less than 1 is due to getting a complete set of SEP event periods but not all of the non-event periods. The impact of not having a complete set of non-event periods is minimal and any comparisons to other SEPVAL models is considered to be without any caveats. We compare MAG4 to SWPC Day-1 forecasts and MagPy in Section 7.1.

MAG4 downloads SDO/HMI full-disk line-of-sight (LOS) and SHARP magnetograms at a one hour cadence and issues a forecast for the probability of a standard definition SPE in a 24-hour prediction window. The magneotgrams are analyzed to identify the neutral line between bipolar active regions. Several magnetic parameters relating to the neutral line are computed by MAG4 that serve as a proxy for magnetic free energy of the active region, related to its eruptive potential. In the model training process, these free energy proxies are computed from a large archive of magnetogram data and used in conjunction with ancillary data on the flaring history of an active region and the SPE events associated with the region to produce a forecast curve. The MAG4 forecast curve relates the magnetogram-based free energy proxies to the frequency of event occurrence. Following the training, MAG4 is able to compute the free energy proxy for a given active region and use the forecast curve to issue the probability that the region will produce an SPE in a 24-hour window.

Variant	Magnetogram	Proxy	Flare History?	Trained On
LOS_FEr	LOS	Gradient	No	MDI
LOS_r	LOS	Gradient	Yes	MDI
SHARP_FE	Vector	Gradient	No	MDI
SHARP	Vector	Shear	Yes	MDI
SHARP_HMI	Vector	Shear	Yes	HMI

Table 6.11: MAG4 variant properties.

Under this scheme, MAG4 is capable of generating several models with varied selections of magnetogram instruments used for training, free energy proxies, and consideration of ancillary data such as previous flaring. Five MAG4 model variants are deployed on the SEP Scoreboard and described in Table 6.11. LOS_FEr, LOS_r, and SHARP_FE use the integrated gradient of the magnetic field along the neutral line as the free energy proxy, while SHARP and SHARP_HMI use the shear along the neutral line. Another key difference between the model types is the source magnetograms: full-disk LOS or vector magnetogram SHARPs. Operationally, the LOS magnetograms have higher availability, as SHARP data require more sophisticated processing to produce, and are occasionally not available at the desired near-real-time cadence. The SHARP vector magnetic field products are a more physically complete representation of solar magnetic fields, and therefore are expected to produce a more accurate free energy proxy for forecasting Falconer et al. (2002).

Due to projection effects of the source magnetogram data, MAG4 probabilities are only reliable when active regions are withing 45 degrees of disk-center. Predictions are still made using regions up to 85 degrees, but accuracy drops significantly. An additional caveat is that MAG4 variants that use active region SHARPs as input

(SHARP, SHARP_FE, SHARP_HMI) will still issue forecasts based on incomplete data when SHARP data is not available from JSOC. These types of data outages are occasional. Such forecasts may omit significant active regions, and as a result will have a reduced event probability. MAG4 re-issues an older forecast in the case that no data is available, however SPHINX removes these forecasts as duplicates. For more rare extended data outages and occasional periods when the source data are corrupted, MAG4 issues a default 1% probability forecast. These undesirable behaviors are folded into the Scoreboard validation analysis, and so the validation should be thought of as representative of observed real-world performance of the full model system, not isolated model performance under idealized conditions when all data is available all the time.

MAG4	LOS_FEr	LOS_r	SHARP_FE	SHARP	SHARP_HMI
Energy	>10 MeV	>10 MeV	>10 MeV	>10 MeV	>10 MeV
N	27137	27150	18699	19009	19288
Brier Score	0.030	0.029	**0.028**	**0.028**	**0.028**
Brier Skill	-0.059	-0.037	-0.036	$\mathbf{-0.032}$	-0.042
Brier (SEP)	**0.88**	0.92	0.91	0.94	0.96
Median P (SEP)	**0.05**	0.03	0.01	0.01	0.01
AUC	**0.65**	**0.65**	0.62	0.61	0.58

Table 6.12: MAG4 probability metrics from Scoreboard forecasts.

MAG4	SEPVAL		Scoreboard	
Variant	LOS_r	SHARP_HMI	LOS_r	SHARP_HMI
Energy	>10 MeV	>10 MeV	>10 MeV	>10 MeV
N	897	547	27150	19288
Brier Score	**0.19**	0.21	0.029	**0.028**
Brier Skill Score	**0.00**	-0.03	$\mathbf{-0.037}$	-0.042
Brier (SEP)	**0.94**	0.97	**0.92**	0.96
Median P (SEP)	**0.03**	0.01	**0.03**	0.01
Area Under the Curve	**0.67**	0.56	**0.65**	0.58

Table 6.13: MAG4 probability metrics from SEPVAL and Scoreboard forecasts.

Table 6.12 shows the probability metrics for all five MAG4 variants running on the Scoreboard, while Table 6.13 shows the SEPVAL results for two selected variants LOS_r and SHARP_HMI alongside the Scoreboard results. Overall, all five MAG4 variants have nearly equivalent performance. The Brier Scores are all small values close to the optimal value of 0, but this is largely a result of the large imbalance of SEP event dataset ($\sim$1 event to every 35 non-events, see Table 6.9) and a large number of low probability forecasts issued by MAG4 entering into the Brier Score average (see definition Equation 3.10). The Brier Skill Scores for all variants are all nearly zero and slightly negative, indicating that the model performance is nearly the same and as the reference model of a constant climatological forecast of 3.3% based on solar cycle 24 statistics (Bain et al., 2021). The negative value indicates

that the model performance is very slightly *worse* than this naive climatological reference model, but one must keep in mind that this is only in a large average sense. The Brier Score and the Brier Skill Score are not effective in showing the models' ability to discriminate between event and non-event periods.

The ability of MAG4 to discriminate events is captured by inspecting the distribution of probability forecasts, as shown in Figure 6.5. Two distributions are shown for each MAG4 variant: forecasts whose 24-hour prediction window contains an SEP event on the top, and those which do not on the bottom. Forecasts matched to a SEP have a probability distribution that is generally broader and with a higher median, indicating a larger proportion of higher probability forecasts preceeding an event. The skill of the model in discriminating events can be evaluated from the median and by the Brier Score for the SEP-matched forecasts. Model variants that do *not* consider previous flaring (FE; free-energy only) outperform those that do for both the line-of-sight (LOS) and Spaceweather HMI Active Region Patch (SHARP) versions. This is a surprising result as it is well known that previous flaring is a robust indicator of future flaring, and flaring and SEP acceleration are related. The issue may come down to a difference or error in the probability scale of the underlying forecast curve. Notice that the cluster of higher-probability forecasts (those at >20%) are higher for the free-energy-only variants than for the flare variants. Both models may have similar ability to distinguish higher-threat active regions, but the free-energy version may be assigning a higher probability per event for those same regions. The best performing variant overall is LOS_FEr, with its higher median probability for SEP-matching forecasts (0.05) and lowest Brier Score for that subsample (Brier Score = 0.88). The discrimination of events is still clearly very low, with the LOS_FEr median probability only rising to 0.05 for SEP-matching forecasts, compared to 0.02 for those matching quiet periods.

Figure 6.6 shows the forecast probability distributions for the two variants participating in SEPVAL: LOS_r and SHARP_HMI. The results for this validation dataset are compatible with what was found for the Scoreboard: the LOS variant outperforms the SHARP variant, showing an increased ability to discriminate SEP event periods from non-event periods. The median probability and Brier Score metrics for the event periods are summarized in Table 6.13.

Figure 6.7 shows the reliability diagrams for the Scoreboard. These diagrams illustrate how accurate the probabilities are, e.g. does a 10% probability forecast precede an event 10% of the time? The bin widths in these diagrams are 10% in probability, the points connected with lines give the reliability, with perfect reliability falling on the 1:1 diagonal line, and the histograms show the number of forecasts in each bin, with a log scale. The Figures show that for LOS_FEr, probabilities are skewed to low for the 10%–30% bins, but are nearly accurate for the 0%–10% bin and for the 30%–50% bins. LOS_r is similarly underestimating for low probabilities, but is somewhat overestimating for higher probabilities. MAG4 SHARP was the most consistently accurate from 0%–30%, but underestimates the highest bin, but statistics are low. All points in the 90%–100% bin are erroneous forecasts likely caused by corrupted input data. MAG4 issued legitimate forecasts only up to 50% probability, putting an upper limit on the certainty of its forecasts.

ROC curves for all MAG4 variants are plotted in Figure 6.8 for the Scoreboard. The skill in discriminating events is also captured by the AUC metric derived from these plots. Here again the scores are very similar across variants, but with the LOS models showing slightly higher skill. This is a surprising result, as the SHARP vector magnetic field data product is a more complete physical representation of the solar magnetic field than the LOS observation, which should allow for the computation of a more accurate free energy proxy. The lower reliability of the real-time data products could also play a factor in the decreased performance, with MAG4 issuing forecasts based on only a subset of the active region (AR)s available at time. However, slightly decreased performance of SHARP_HMI versus LOS_r is also seen in the SEPVAL results, where real-time data availability is not an issue. See Figure 6.9 for ROC curves from SEPVAL. A more detailed analysis is required to determine the reasons for the performance difference between LOS and SHARP variants. One important consideration is the possibility of biases in the SEPVAL non-event dataset. While the Scoreboard non-events are representative of the real Sun climatology, the SEPVAL non-events were selected to have flare & CME characteristics similar to the event set. This validation set may be a particularly challenging set for MAG4, highlighting its ability to distinguish regions that produce SEPs against regions that merely produce notable flares & CMEs but without SEPs.

MAG4	LOS_FEr	LOS_r	SHARP_FE	SHARP	SHARP_HMI
Energy	> 10 MeV	> 10 MeV	> 10 MeV	> 10 MeV	> 10 MeV
N	1271	1275	1233	1233	1233
Hits	29	28	23	24	13
False Alarms	848	735	438	549	183
Correct Negatives	388	505	761	650	1016
Misses	6	7	11	10	21

Table 6.14: MAG4 All Clear contingency table, deoverlapped Scoreboard.
MAG4 All Clear contingency table values from deoverlapped Scoreboard forecasts.

MAG4	LOS_FEr	LOS_r	SHARP_FE	SHARP	SHARP_HMI
Energy	> 10 MeV	> 10 MeV	> 10 MeV	> 10 MeV	> 10 MeV
N	1271	1275	1233	1233	1233
Percent Correct	0.33	0.42	0.55	0.64	**0.83**
Hit Rate	**0.83**	0.80	0.71	0.68	0.38
False Alarm Rate	0.69	0.59	0.46	0.37	**0.15**
False Alarm Ratio	0.97	0.96	0.96	0.95	**0.93**
Threat Score	0.03	0.04	0.04	0.05	**0.06**
Bias	25.1	21.8	16.9	13.6	**5.8**
HSS	0.01	0.02	0.03	0.04	**0.07**
TSS	0.14	0.21	0.25	**0.31**	0.23

Table 6.15: MAG4 All Clear metrics from deoverlapped Scoreboard forecasts.

MAG4 produces an All Clear forecast by thresholding its probability forecast. The chosen threshold was selected to be 1%, the minimum probability that can be issued by MAG4. All forecasts exceeding 1% are "Not Clear". The selection of such

SEPVAL

MAG4_LOS_r		Observed Yes	Observed No	Sum
	Pred. Yes	22	14	36
	Pred. No	11	9	20
	Sum	33	25	56

MAG4_SHARP_HMI		Observed Yes	Observed No	Sum
	Pred. Yes	16	8	24
	Pred. No	17	16	33
	Sum	33	24	57

Scoreboard

		Observed Yes	Observed No	Sum
	Pred. Yes	28	735	763
	Pred. No	7	505	512
	Sum	35	1240	1275

		Observed Yes	Observed No	Sum
	Pred. Yes	13	183	196
	Pred. No	21	1016	1037
	Sum	34	1199	1233

Table 6.16: Contingency tables for MAG4 variants common to both SEPVAL and the SEP Scoreboards.

MAG4	SEPVAL		Scoreboard	
Variant	LOS_r	SHARP_HMI	LOS_r	SHARP_HMI
Energy	> 10 MeV	> 10 MeV	> 10 MeV	> 10 MeV
N	57	56	1275	1233
Percent Correct	0.55	**0.56**	0.42	**0.83**
Hit Rate	**0.67**	0.48	**0.80**	0.38
False Alarm Rate	0.61	**0.33**	0.59	**0.15**
False Alarm Ratio	0.39	**0.33**	0.96	**0.93**
Threat Score	**0.47**	0.39	0.04	**0.06**
Bias	**1.09**	0.73	21.8	**5.8**
HSS	0.06	**0.14**	0.02	**0.07**
TSS	0.06	**0.15**	0.21	**0.23**

Table 6.17: MAG4 All Clear metrics from SEPVAL and Scoreboard forecasts.

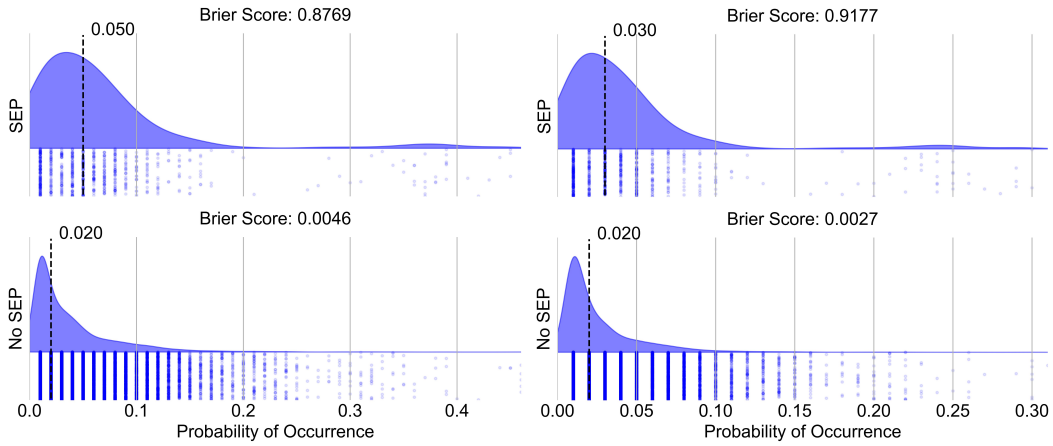
a low threshold biases the model towards a high Hit Rate but with a large number of false alarms, and this can be seen in the All Clear metrics shown in Figure 6.15 for the Scoreboard and Table 6.17 for SEPVAL compared to the Scoreboard. Note that we calculate All Clear metrics from deoverlapped (see Section 3.5) MAG4 forecasts. Interestingly, at this threshold our ranking of model performance is largely the opposite of what is seen in the probability forecast analysis above: LOS_FEr has the lowest performance in terms of the skill score HSS and TSS, while SHARP_HMI has the highest performance in terms of HSS and SHARP has the highest TSS. In general, the flare-aware variants have higher performance than the flare-ignorant, and the SHARP variants have higher performance than the LOS counterparts. Objectively, however, all variants have an extremely high number of false alarms, with False Alarm Ratios exceeding 90%, indicating that the vast majority of all "Not Clear" forecasts turn out to be clear. Hit Rates are high, however for all but the

SHARP_HMI variant: with 68% to 83% of event days having a "Not Clear" forecast. The SHARP_HMI variant has only a 38% Hit Rate, but a correspondingly lower false alarm rate and ratio pushes up its skill scores (HSS and TSS) in this highly unbalanced dataset (1 event days for every 35 non-event days).

MAG4's selection of the lowest possible All Clear probability threshold results in an All Clear performance that maximizes the Hit Rate at the expense of a large number of false alarms. This leads to very poor skill overall, and inspection of the ROC curves in Figures 6.8 and 6.9 indicate that larger skill scores are possible with the selection of a higher probability threshold. Taking LOS_FEr as an example, the 1% probability threshold corresponds to the point (0.69, 0.83) on the corresponding ROC curve. A more optimal All Clear skill is likely attainable by choosing the probability threshold corresponding to the (0.4, 0.65) point. This would correspond to a Hit Rate of 65%, approximately 20% lower than from using the current threshold, but a ∼30% lower False Alarm Rate. This tradeoff would likely result in higher skill scores, however more analysis would be required to find the optimal probability threshold.

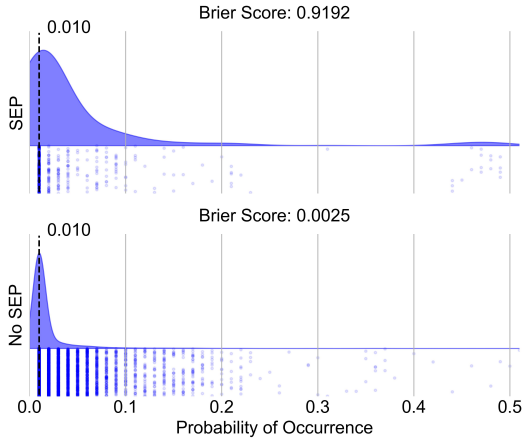
The main takeaways from this validation are summarized below:

- The performance of the MAG4 probability forecasts are nearly equivalent across all five model variants, with variants based on LOS observations slightly outperforming SHARP-based variants in discriminating events from non-events.
- Free-energy-only variants outperform those that consider previous flaring in discriminating event periods from non-event periods in terms of probability. The highest performing variant overall is LOS_FEr.
- The ability of MAG4 to discriminate event and non-event periods in terms of probability is objectively quite low. For the best performing variant (LOS_FEr), forecasts that preceed SEP events have only a slightly higher (+4%) increase in their median probability over those that cover quiet periods.
- The selection of a 1% probability threshold to discriminate All Clear results in poor model skill, with high Hit Rates from 68%–83% depending on the variant, but very large False Alarm Ratios exceeding 95%. The corresponding HSS scores are near 0 indicating performance similar to that of a random forecast model, and TSS scores are also low, ranging from 0.14 to 0.31.
- The selection of a 1% probability threshold for All Clear inverts the sense of "best" model with respect to what was described above for probability forecasts, with SHARP variants outperforming LOS and flare-aware variants outperforming flare-ignorant ones.
- Inspection of the ROC curves indicates higher All Clear performance could be achieved with the selection of a higher probability threshold for LOS variants.

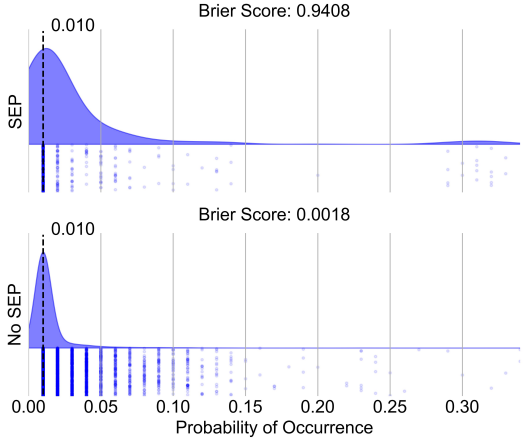


(a) MAG4_LOS_FEr

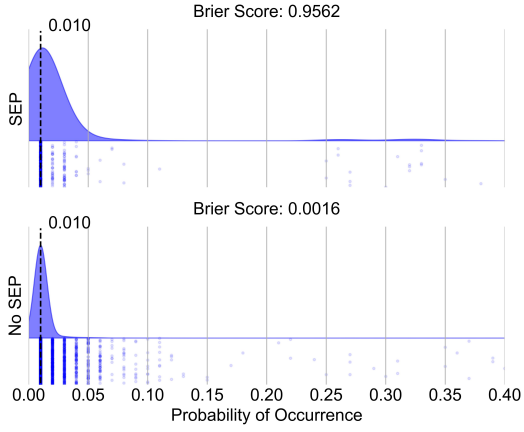
(b) MAG4_LOS_r



(c) MAG4_SHARP_FE

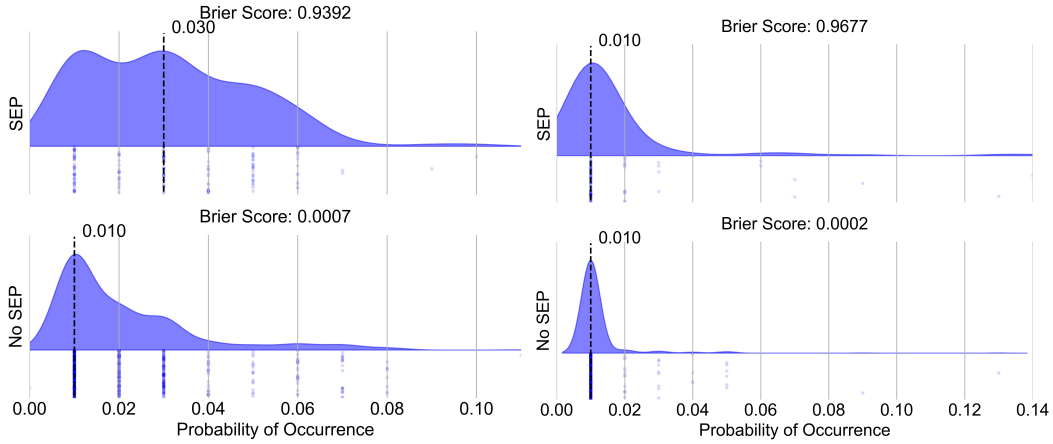


(d) MAG4_SHARP



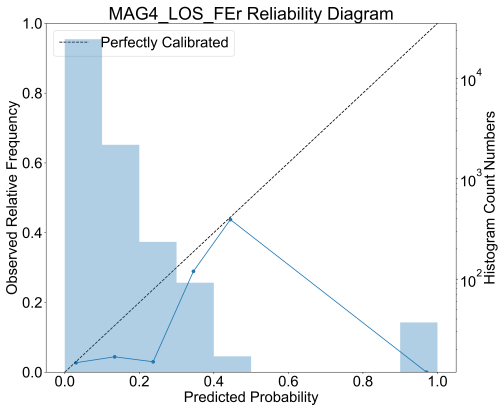
(e) MAG4_SHARP_HMI

Figure 6.5: Forecasted probabilities and Gaussian kernel density estimates for MAG4 variants from SEP Scoreboard, separated by event periods and non-event periods. Considers >10 MeV integral proton flux with a 10 pfu flux threshold.

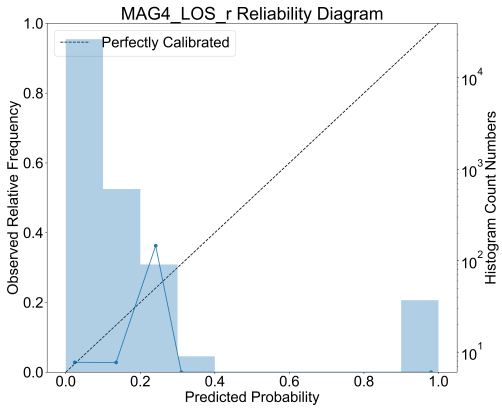


(a) MAG4_LOS_r (b) MAG4_SHARP_HMI

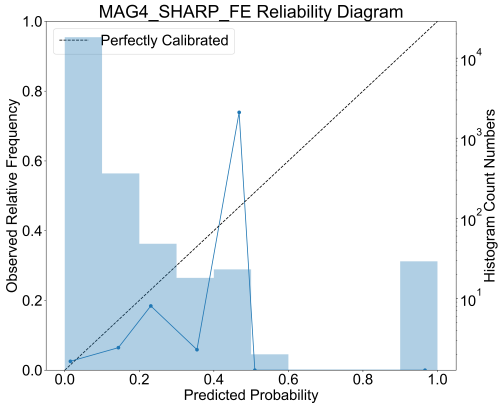
Figure 6.6: Forecasted probabilities and Gaussian kernel density estimates for MAG4 variants from SEPVAL, separated by event periods and non-event periods. Considers >10 MeV integral proton flux with a 10 pfu flux threshold.



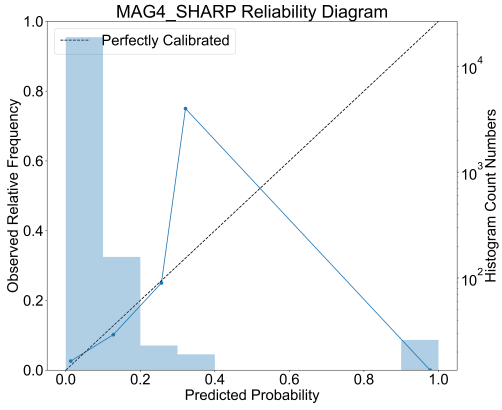
(a) MAG4_LOS_FEr



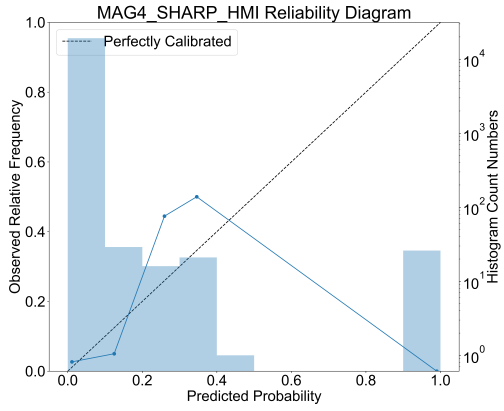
(b) MAG4_LOS_r



(c) MAG4_SHARP_FE

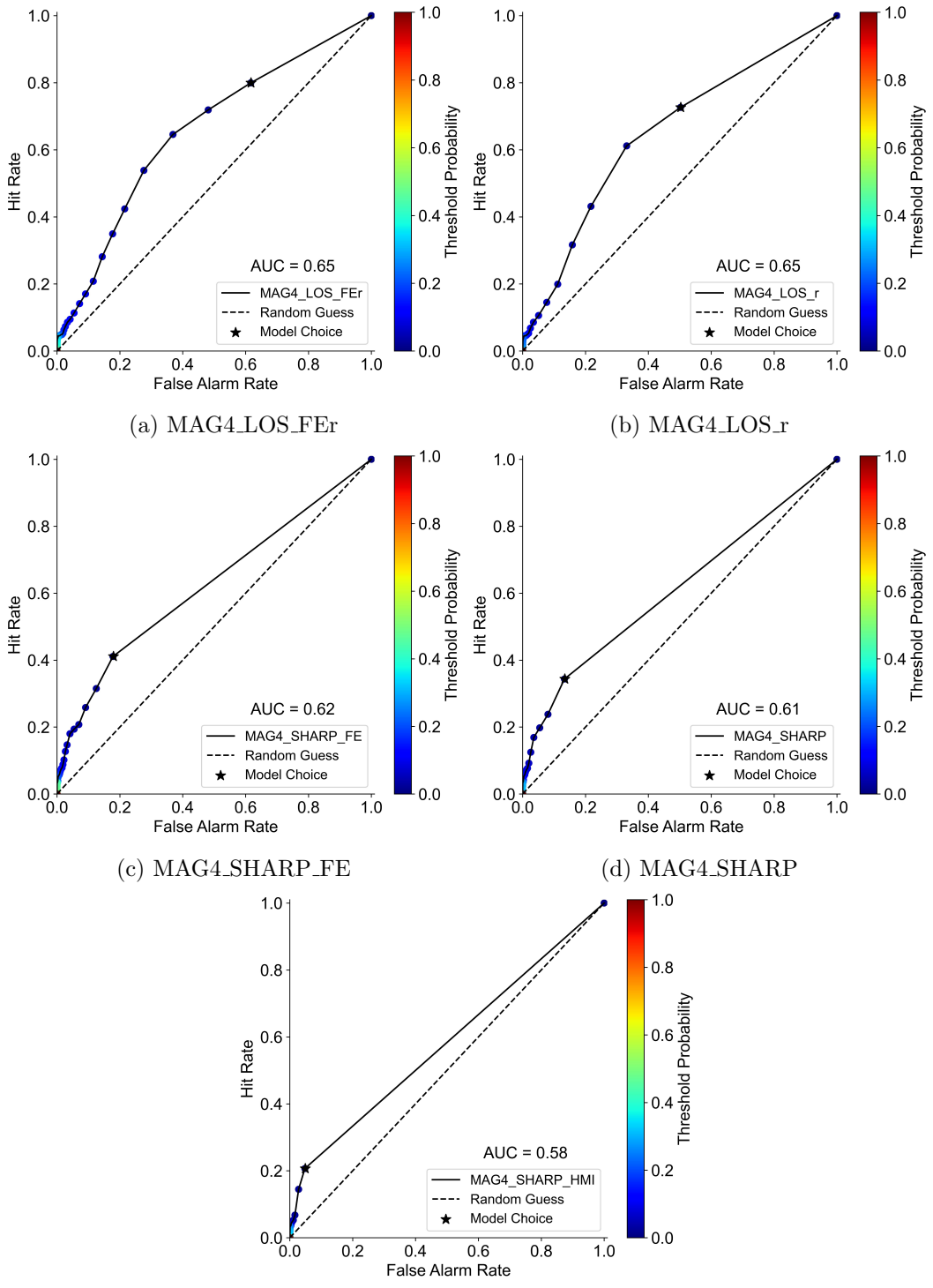


(d) MAG4_SHARP



(e) MAG4_SHARP_HMI

Figure 6.7: Reliability diagrams for MAG4 variants from SEP Scoreboard. Considers >10 MeV integral proton flux with a 10 pfu flux threshold.



(a) MAG4_LOS_FEr

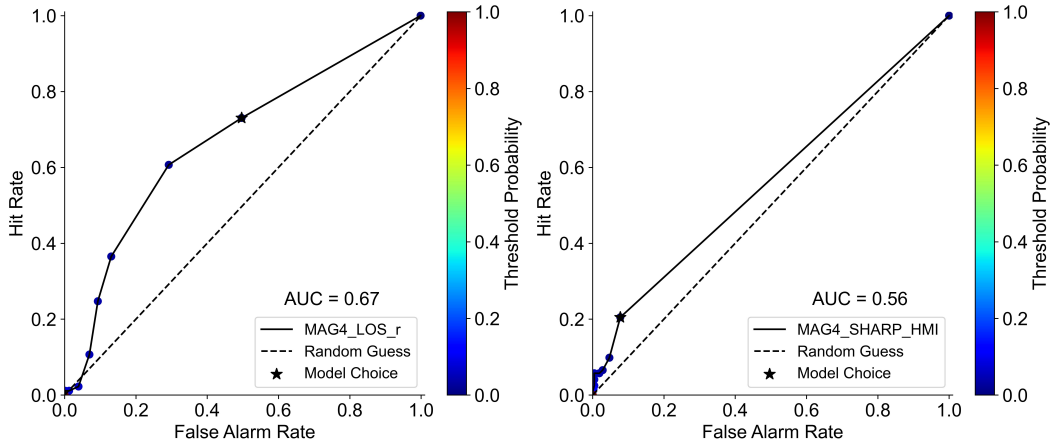
(b) MAG4_LOS_r

(c) MAG4_SHARP_FE

(d) MAG4_SHARP

(e) MAG4_SHARP_HMI

Figure 6.8: ROC curves for MAG4 variants from SEP Scoreboard. Considers >10 MeV integral proton flux with a 10 pfu flux threshold.



(a) MAG4_LOS_r (b) MAG4_SHARP_HMI

Figure 6.9: ROC curves for MAG4 variants from SEP SEPVAL. Considers >10 MeV integral proton flux with a 10 pfu flux threshold.

6.3 MagPy

MagPy is the Python-based evolution of MAG4, designed as a near real-time forecasting tool for predicting an active region's next-day production rate of major flares, CMEs, and SEP events. While MAG4 was originally developed in IDL, its platform presents limitations for long-term development. The MagPy project began as a straightforward port of MAG4 to the Python language, but eventually incorporated optimizations and extensions upon the MAG4 paradigm that now make it a distinct model in its own right.

MagPy SHARP HMI	
Characteristic	> 10 MeV
First Forecast	2023-06-27
Last Forecast	2024-11-27
N Days	520
N Forecast Days	517
N SEP Days	23
Forecast Cadence	1 hr
Prediction Window	24 hr
N forecasts	11,849
N matched w/events	525
Imbalance (raw)	21.6
Imbalance (days)	21.5

Table 6.18: MagPy SHARP HMI validation characteristics.

Like MAG4, MagPy takes real-time magnetograms as an input and calculates proxies for magnetic free energy based on either the shear or gradient of strong fields along the polarity inversion line. The SHARP_HMI_CEA variant of MagPy analyzed here uses SHARPs downloaded from the JSOC to forecast SEP events using the magnetic shear, and is analogous to the MAG4 variants SHARP and SHARP_HMI. MagPy SHARP_HMI_CEA began producing forecasts on the Scoreboard in June of 2023, and has produced over 11,000 forecasts through the end of 2024 (see Table 6.18). At the time of writing, a LOS variant of MagPy has been completed, but has not yet been deployed on the Scoreboard. A complete verification, validation, and comparison of MagPy LOS to MAG4 will be done in a future work.

In order to optimize performance over MAG4, MagPy underwent an extensive training and testing cycle, processing years of historical data to refine the relationship between an AR's total non-potentiality and its likelihood of producing major solar flares or SEPs. By testing thousands of magnetic field threshold configurations and comparing skill scores across validation results, the most effective set of operational thresholds was determined. This optimization effort showed significantly improved forecasting accuracy, increasing the HSS from 0.32 in MAG4 to 0.48 in MagPy based on the internal model development dataset. An optimization of the probability threshold for declaring All Clear was also done. A threshold of 19% was

found to optimize the HSS on the development dataset. This is notably higher than the 1% threshold used for MAG4 All Clear. Based on feedback from the SEPVAL validation, MagPy's All Clear probability threshold was reduced to 15% in September 2024. Approximately two months of MagPy Scoreboard forecasts ($\sim$ 10%) have this lower probability threshold.

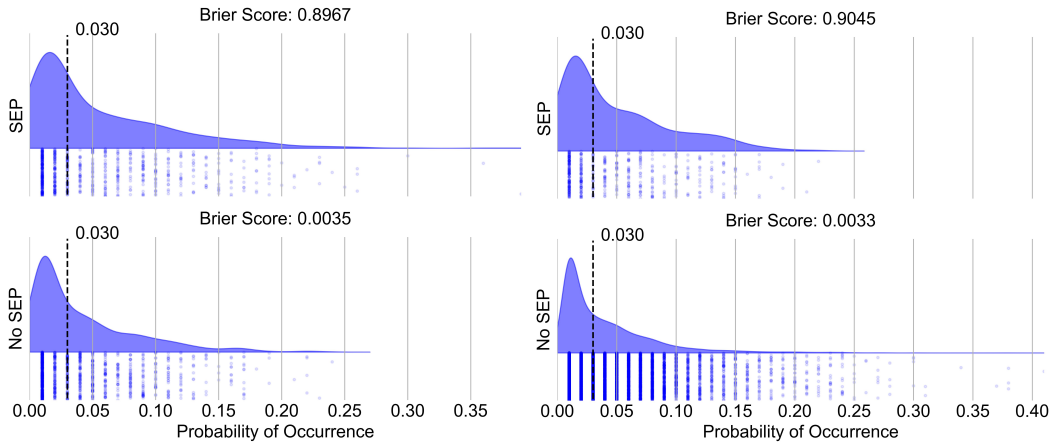
Note that the HSS seen in MagPy development are not expected in our validation, as this metric is extremely sensitive to the imbalance in the underlying dataset. However the improvement over MAG4 that was seen in the development process is also expected in our validation, but a fair comparison can only be done when both models are evaluated on similar datasets. The performance of MagPy is compared to MAG4 in Section 7.1. The present section only considers MagPy performance in isolation.

Like MAG4, MagPy's free energy proxy calculations are increasingly susceptible to errors due to projection effects as active regions approach the solar limb. Active regions within a 45-degree cone from disk center are the most reliable, and in the training procedure only active regions within 45 degrees are considered. However, as an additional improvement over MAG4, an empirical analysis was done to find a compensation curve that provides a weighting to apply to the free-energy proxies for active regions on the limb in order to restore their value to what would be expected if the region were near disk center. This limb correction is applied in MagPy during real-time operation, allowing it to more reliably use observations out to 85 degrees from disk center. Unfortunately, this limb correction was not well-calibrated in the early deployment of MagPy, leading to systematically higher probability forecasts as major active regions reached the limb. This systematic error is still present in the Scoreboard forecasts and is incorporated into our validation analysis, likely negatively impacting its performance results to some degree. The issue was corrected and the SEPVAL challenge set forecasts were computed with a version of MagPy using the re-calibrated limb correction.

MagPy SHARP HMI Probability	SEPVAL > 10 MeV ($N = 2003$)	Scoreboard > 10 MeV ($N = 11117$)
Brier Score	0.36	0.05
Brier Skill Score	0.04	-0.01
Brier (SEP)	0.896	0.905
Median P SEP	0.03	0.03
Area Under the Curve	0.56	0.56

Table 6.19: MagPy SHARP HMI probability metrics.

Table 6.19 gives the probability metrics for MagPy SHARP_HMI, while Figure 6.10 shows the distributions of the probability forecasts. Metrics and distributions for both datasets are very similar, with the exception of the Brier Score which significantly decreases from 0.36 to 0.05 due to the greatly increased number of non-event forecast periods in the Scoreboard dataset. MagPy's ability to distinguish event periods from non-event periods is extremely limited, as seen in Figure 6.10.



(a) MagPy SEPVAL

(b) MagPy Scoreboard

Figure 6.10: Forecasted probabilities and Gaussian kernel density estimates for MagPy from SEPVAL and the SEP Scoreboard, separated by event periods and non-event periods. Considers >10 MeV integral proton flux with a 10 pfu flux threshold.

While the long tail of higher probability forecasts is fatter for SEP periods, the median value is unchanged and remains a low 3%. For the Scoreboard dataset, the high-probability tail of the distribution is longer for the non-event periods, which is the opposite of what is desired. This may be a consequence of the systematic error in the limb-correction curve described above, as this feature was not seen in the SEPVAL dataset where that issue was not present.

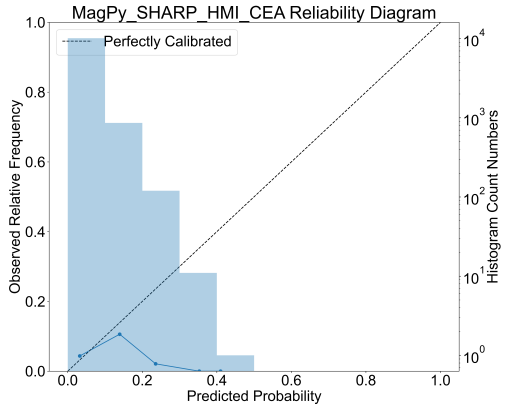
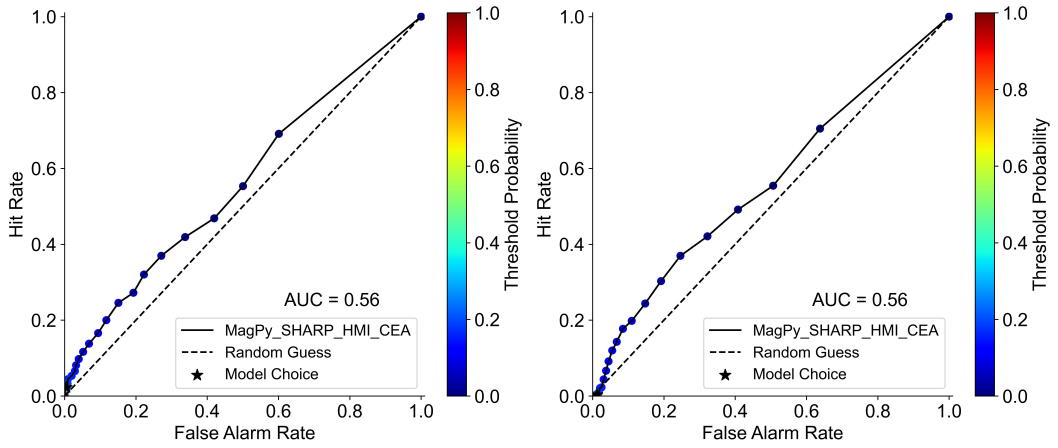


Figure 6.11: Reliability diagram for MagPy SHARP HMI CEA from the SEP Scoreboard. Considers >10 MeV integral proton flux with a 10 pfu flux threshold.

Figure 6.11 shows the reliability diagram for MagPy on the Scoreboard. We see here that probability forecasts in the range 1%–20% are accurate – the observed

event frequency closely matches the predicted probability. However, for higher probability forecasts the observed frequency is substantially below the prediction. This is another reflection of the differences in the high probability (> 20%) tails seen in Figure 6.10, with the non-event distribution having a longer tail than the event distribution, indicating probability "misses". MagPy did not issue probability forecasts in excess of 40%.



(a) MagPy SEPVAL

(b) MagPy Scoreboard

Figure 6.12: ROC curves for MagPy SHARP HMI CEA from SEPVAL and the SEP Scoreboard. Considers >10 MeV integral proton flux with a 10 pfu flux threshold.

		SEPVAL			Scoreboard		
		Observed			Observed		
		Yes	No	Sum	Yes	No	Sum
MagPy SHARP HMI 19%	Pred. Yes	4	2	6	2	26	28
	Pred. No	29	27	56	21	457	478
	Sum	33	29	62	23	483	506

		Observed		
		Yes	No	Sum
MagPy SHARP HMI 7%	Pred. Yes	22	17	39
	Pred. No	11	12	23
	Sum	33	29	62

Table 6.20: Contingency tables for MagPy SHARP HMI variants common to both SEPVAL and the SEP Scoreboards with deoverlapping applied.

Figure 6.12 shows the ROC curves for MagPy for both SEPVAL and the Scoreboard datasets. The indicated performance is very similar across both datasets, with an AUC metric of 0.56 for both. The performance is not substantially above

MagPy SHARP HMI	SEPVAL		Scoreboard
All Clear	> 10 MeV	> 10 MeV	> 10 MeV
Probability Threshold	19%	7%	19%*
N	62	62	506
Percent Correct	0.50	**0.55**	0.91
Hit Rate	0.12	**0.67**	0.09
False Alarm Rate	**0.07**	0.59	0.05
False Alarm Ratio	**0.33**	0.44	0.91
Bias	1.18	1.18	1.22
Threat Score	0.11	**0.44**	0.04
HSS	0.05	**0.08**	0.03
TSS	0.05	**0.08**	0.03

Table 6.21: MagPy SHARP HMI All Clear Metrics from SEPVAL and the SEP Scoreboard. Best values in the SEPVAL comparison are in boldface. *Note that the Scoreboard probability threshold changed from 19% to 15% late in the validation period.

that of a random model indicated by the diagonal line.

Table 6.21 shows the All Clear metrics for MagPy with deoverlapping applied. The high probability threshold of 19% combined with a low numbers of such probabilities during SEP events (Figure 6.10) results in extremely low Hit Rates for both the SEPVAL and the Scoreboard (12% and 9%, respectively), and correspondingly low skill in HSS and TSS. Despite this, the False Alarm Ratio in real-time operation is still very high, 91%. All Clear performance in real-time operations could be somewhat improved by the selection of a lower probability threshold. Figure 6.12 shows the Hit Rate and False Alarm Rate for a series of integer probability thresholds ranging from 0% on the top-right (1.0, 1.0), and proceeding down to 100% on the bottom left (0.0, 0.0). By counting the points down the curve, we can see that a choice of 9% probability would result in a Hit Rate of ∼20% and a False Alarm Rate of ∼10%. All Clear metrics on the SEPVAL dataset were recomputed for a series of probability thresholds, and a value of 7% was found to be optimal. SEPVAL results for the 7% threshold are shown in Table 6.21. It is seen that this choice results in a substantial increase in the Hit Rate (from 12% to 67%) at the cost of a similar increase in the False Alarm Rate (from 0.07% to 0.59%). The skill scores HSS and TSS receive a modest increase, from 0.05 to 0.08. The tradeoff is stark, and the best choice depends on what is considered more valuable, hits or correct negatives. Improvements beyond these require more fundamental improvement to the underlying probability forecasts placing points closer to the optimal (0.0, 1.0) corner of the ROC curve.

The main takeaways from the validation of MagPy are listed below:

- The ability of MagPy SHARP_HMI_CEA to discriminate event and non-event periods is very poor. The median probability forecast is nearly unchanged at 3% for both SEP and non-SEP periods for both SEPVAL and the SEP

Scoreboard datasets.

- Higher probability forecasts do appear to be relatively more frequent during SEP periods, but we have not yet developed a quantitative metric that captures the extent of this feature.
- MagPy probability forecasts are reliable in the 0–20% range, but forecasts in the 20%–40% range overestimate the likelihood of an event. MagPy does not issue probability forecasts exceeding 40%.
- MagPy's All Clear performance is heavily weighted towards correctly predicting non-event periods, with a low False Alarm Rate of 5% corresponding to 95% of such periods forecasted correctly. But the model has low All Clear skill, with a low Hit Rate of 9% and HSS & TSS at 0.03.
- MagPy could increase its All Clear skill with a lower probability threshold, at the expense of increasing its False Alarm Rate. Changing the probability threshold from 17% to 9% would result in a Hit Rate of ∼20% and a False Alarm Rate of ∼10%.

6.4 GSU All Clear

The GSU All Clear model was the result of a 3-year contract with SRAG to develop an advanced All Clear model using machine-learning techniques and SHARP magnetograms as an input. SHARPs are used in two ways: (1) as direct inputs to deep convolutional neural networks, and (2) parameters are extracted from the data to produce time series of active region evolution that are fed into separate machine learning models. The forecasting system consists of a series of models that feed forecasts downstream to predict All Clear, beginning with a flare predictor from the active region, a CME predictor from the flare, and finally an SEP prediction from the CME. The model was deployed on the Scoreboard in April of 2023 and has produced over 12,000 forecasts, 560 of them matched with SEP events (see Table 6.22. The model did not participate in the SEPVAL community validation challenge.

GSU All Clear	
Characteristic	> 10 MeV
First Forecast	2023-04-10
Last Forecast	2024-11-27
N Days	598
N Forecast Days	561
N SEP Days	24
Forecast Cadence	1 hr
Prediction Window	24 hr
N forecasts	12,857
N matched w/events	560
Imbalance (raw)	22.0
Imbalance (days)	22.3

Table 6.22: GSU All Clear validation characteristics.

GSU All Clear Probability	SEPVAL > 10 MeV ($N = -$)	Scoreboard > 10 MeV ($N = 12077$)
Brier Score	-	0.27
Brier Skill Score	-	-5.19
Brier (SEP)	-	0.23
Median P SEP	-	0.54
Area Under the Curve	-	0.53

Table 6.23: GSU All Clear probability metrics.

Table 6.23 gives the probability metrics for the GSU All Clear model, while Figure 6.13 shows the distribution of probability forecasts for event and non-event periods. The probability distributions are puzzling. For non-event periods the forecasted probabilities appear to be a normal distribution centered at 0.51 (median value). The model issues event probabilities in a wide range from $\sim 20\%$ to $\sim 80\%$,

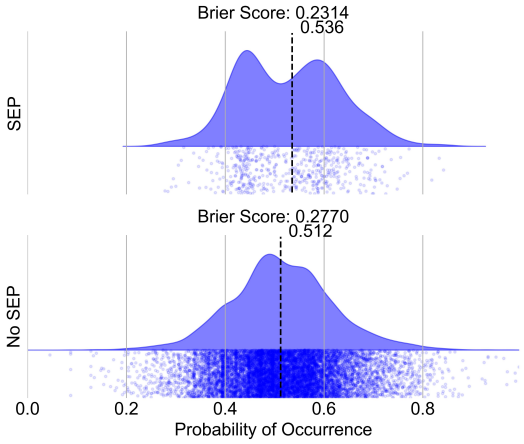


Figure 6.13: Forecasted probabilities and Gaussian kernel density estimates for GSU All Clear from the SEP Scoreboard, separated by event periods and non-event periods. Considers >10 MeV integral proton flux with a 10 pfu flux threshold.

infrequently issuing low probability forecasts that are expected to be common for such rare events. During event periods, the distribution of probability forecasts is bimodal, with nearly equal magnitude peaks at $\sim$ 45% and $\sim$ 60%, resulting in a median probability of 53.6%. The result is very poor model performance, far from the ideal behavior of low probability forecasts during non-event periods and high probability forecasts for event periods. As a result, the performance metrics in Table 6.23 are extremely poor, with the Brier Skill Score reaching a large negative value of -5.19 and the area under the curve metric 0.53 falling very close to that of a random guess model (0.5).

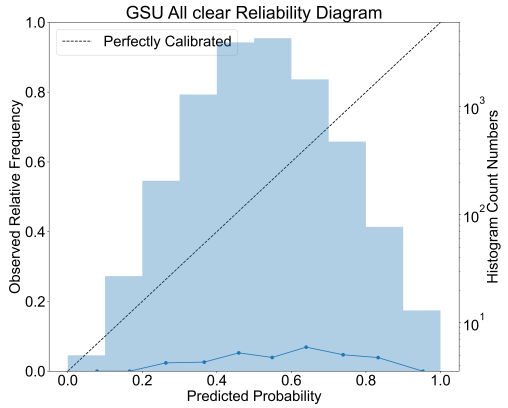


Figure 6.14: Reliability diagram for GSU All Clear from the SEP Scoreboard. Considers >10 MeV integral proton flux with a 10 pfu flux threshold.

Indeed, the reliability diagram shown in Figure 6.14 is similar to that of a random model with a mean probability of 0.5. There appears to be no relationship between

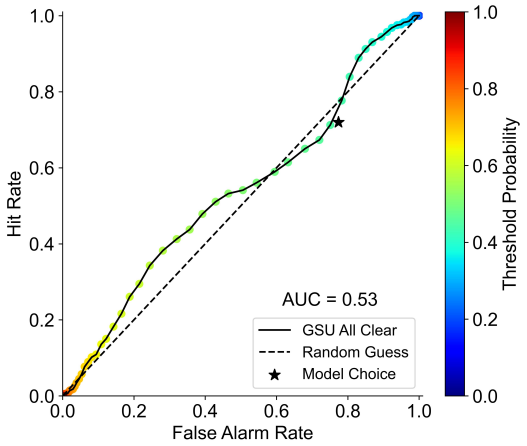


Figure 6.15: ROC curve for GSU All Clear from the SEP Scoreboard, separated by event periods and non-event periods. Considers >10 MeV integral proton flux with a 10 pfu flux threshold.

the predicted probability and the observed frequency of events. The ROC curve shown in Figure 6.15 wavers about the diagonal line of a random guess model.

	Observed		
	Yes	No	Sum
Pred. Yes	23	481	504
Pred. No	1	48	49
Sum	24	529	553

Table 6.24: GSU All Clear contingency table for the SEP Scoreboard with deoverlapping applied.

GSU All Clear All Clear	SEPVAL > 10 MeV $(N = -)$	Scoreboard > 10 MeV $(N = 553)$
Percent Correct	-	0.13
Hit Rate	-	0.96
False Alarm Rate	-	0.91
False Alarm Ratio	-	0.95
Bias	-	21.0
Threat Score	-	0.05
HSS	-	0.00
TSS	-	0.05

Table 6.25: GSU All Clear metrics.

Table 6.25 gives the All Clear binary forecast performance for the GSU All Clear model. The All Clear forecast is based on a probability threshold, and as the

underlying probability model is nearly indistinguishable from a random model, the resulting All Clear performance is similarly poor. Indeed, the HSS metric, designed to indicate model skill above that of a random model, is exactly 0.00. The high hit rate of 96% can be understood as an artifact of our deoverlapping procedure (see Section 3.5). Any day (24-hour period) that contains a "Not Clear" forecast is considered "Not Clear" for that day. Since the model is essentially random and contains 24 hourly forecasts per day, nearly every day is "Not Clear", resulting in nearly every event being "Hit".

In conclusion, the model performance can be summarized as follows:

- The GSU All Clear model demonstrates no skill whatsoever in forecasting SEP events. The model is nearly indistinguishable from a random number generator drawing from a Gaussian distribution centered at 50% probability. The HSS of 0.00 accurately reflects this.
- Curiously, the distribution of probability forecasts issued during event periods is bimodal and centered at 53% probability, distinctly different than the distribution for non-event periods. The model appears to be responding to observations made in advance of SEP events, but not in a way that results in an improved forecast.

6.5 SPRINTS

SPRINTS is an empirical and machine-learning forecasting tool developed by NextGen Federal Systems. It was originally developed from the MAG4 model, extending forecast windows out to 72 hours, as a pre-eruptive probability model. The version that is running now on the SEP Scoreboards, as well as for the SEPVAL challenge, is post-eruptive, triggering off of GOES X-rays to provide a forecast once a flare has occurred, however it is implemented somewhat differently in the two cases. For SEPVAL, SPRINTS submitted forecasts using the approach initially developed for the post-eruptive model concept, which generates forecasts starting at the flare start time and updates the forecast every minute until the flare end time (defined as a certain percentage of the peak value). On the SEP Scoreboards, SPRINTS produces only one forecast per flare, waiting until the flare has reached its end to issue a forecast, the equivalent of the last forecast in the series. SPRINTS has multiple prediction windows from 0–24 hours, 24–48 hours, 48–72 hours, and 72–96 hours. Only the 0–24 hour prediction window is validated for this report. The statistics for this model on the SEP Scoreboards are in Table 6.26.

	SPRINTS	
Characteristic	$>$ 10 MeV	$>$ 100 MeV
First Forecast	2022-08-25	2022-08-25
Last Forecast	2024-11-26	2024-11-26
N Days	826	826
N Forecast Days	806	806
N SEP Days	30	5
Forecast Cadence	Triggered	Triggered
Prediction Window	24 hours	24 hours
N forecasts	7519	7519
N matched w/events	38	7
Imbalance (raw)	196.9	1073.1
Imbalance (days)	25.9	160.2

Table 6.26: SPRINTS validation characteristics.

Since SPRINTS gives a probability of the occurrence of an SEP event following a X-ray flare trigger, the conversion to a binary All Clear forecast is determined by applying a probability threshold of 68% for SEPVAL and a threshold of 56% on the SEP Scoreboards. The contingency tables for SEPVAL and the SEP Scoreboards are in Table 6.27 and the metrics are in Table 6.28.

For SEPVAL, SPRINTS performs similarly for both >10 and >100 MeV channels with moderate skill in determining All Clear. Percent Correct for >10 MeV (>100 MeV) is 0.70 (0.67) and Hit Rate is 0.67 (0.60). For the >10 MeV protons, the moderately low False Alarm Rate of 0.27 and False Alarm Ratio of 0.29 show no reliance on overprediction to achieve a high Hit Rate. However, the >100 MeV probabilities show more false alarms with a False Alarm Rate of 0.31 and False Alarm Ratio of 0.61. Moderate values in Threat Score (0.53) and TSS (0.40) for the

SEPVAL

> 10 MeV

	Observed		
	Yes	No	Sum
Pred. Yes	20	8	28
Pred. No	10	22	32
Sum	30	30	60

> 100 MeV

	Observed		
	Yes	No	Sum
Pred. Yes	9	14	23
Pred. No	6	31	37
Sum	15	45	60

Scoreboard

> 10 MeV

	Observed		
	Yes	No	Sum
Pred. Yes	5	14	19
Pred. No	33	7140	7173
Sum	38	7154	7192

> 100 MeV

	Observed		
	Yes	No	Sum
Pred. Yes	3	21	24
Pred. No	4	7467	7471
Sum	7	7488	7495

Table 6.27: SPRINTS contingency tables for SEPVAL and the SEP Scoreboards.

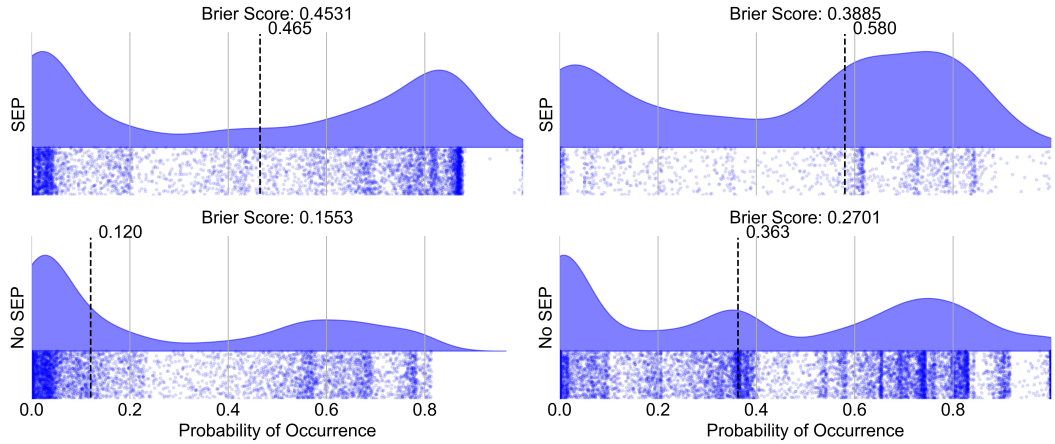
>10 MeV channel, mean that the model differentiates between the conditions that cause SPE events and those that do not with some skill. The >100 MeV forecasts have a more difficult time differentiating between the conditions causing ESPE periods, with Threat Score of 0.31 and TSS of 0.29. HSS shows the skill of the model compared to random chance, where a 0 would be equivalent to random, and for the balanced dataset of SEPVAL, SPRINTS shows skill better than random with HSS = 0.40 for >10 MeV. For >100 MeV, SPRINTS achieves a lower value of HSS = 0.29.

On the SEP Scoreboards, SPRINTS All Clear forecasts show lower skill in the context of the true, highly unbalanced climatology of SEP events. The high Percent Correct of 0.993 and 0.997 for >10 MeV and >100 MeV respectively, are due to the imbalance of the dataset favoring correct negatives. A Hit Rate of 0.13 for the >10 MeV shows that the model misses most SPEs, while the False Alarm Ratio of 0.73 shows that most "yes" forecasts are false alarms. The >100 MeV forecasts have a higher Hit Rate of 0.43 but also a higher False Alarm Ratio of 0.88. The widely differing performance of SPRINTS seen in SEPVAL vs. the Scoreboard may be due to differences in model implementation or reflective of the challenges of forecasting in real time.

Since SPRINTS is a probabilistic model, we can use the next set of metrics, seen in Table 6.29 to determine if the probabilities that SPRINTS provides are reflective of changing solar conditions that are related to SEP events. Brier Score is a measure of probabilistic error, meaning lower scores are favorable. This measure is biased for unbalanced datasets, where consistently issuing a low probability when events are rare (like for SEP events) results in a low Brier Score. SPRINTS achieves low Brier Scores for the SEP Scoreboards due to this reason, with 0.0058 and 0.0018 for >10 MeV and >100 MeV, respectively. However, for the balanced dataset of SEPVAL, there is more meaning behind this metric where the model does show some skill with 0.30 for >10 MeV and 0.29 for >100 MeV. The Brier Skill Score measures model skill relative to climatology. As mentioned in Section 3.3.2, the reference value used was the Solar Cycle 24 average SEP probability provided in (Bain et al., 2021). This

SPRINTS All Clear	SEPVAL > 10 MeV ($N = 60$)	SEPVAL > 100 MeV ($N = 60$)	Scoreboard > 10 MeV ($N = 7192$)	Scoreboard > 100 MeV ($N = 7495$)
Percent Correct	0.70	0.67	0.993	0.997
Hit Rate	0.67	0.60	0.13	0.43
False Alarm Rate	0.27	0.31	0.001	0.003
False Alarm Ratio	0.29	0.61	0.73	0.88
Bias	0.93	1.53	0.50	3.25
Threat Score	0.53	0.31	0.10	0.11
HSS	0.40	0.25	0.17	0.19
TSS	0.40	0.29	0.13	0.43

Table 6.28: SPRINTS All Clear Metrics.



(a) >10 MeV

(b) >100 MeV

Figure 6.16: Forecasted probabilities and Gaussian kernel density estimates for SPRINTS from SEPVAL, separated by event periods and non-event periods

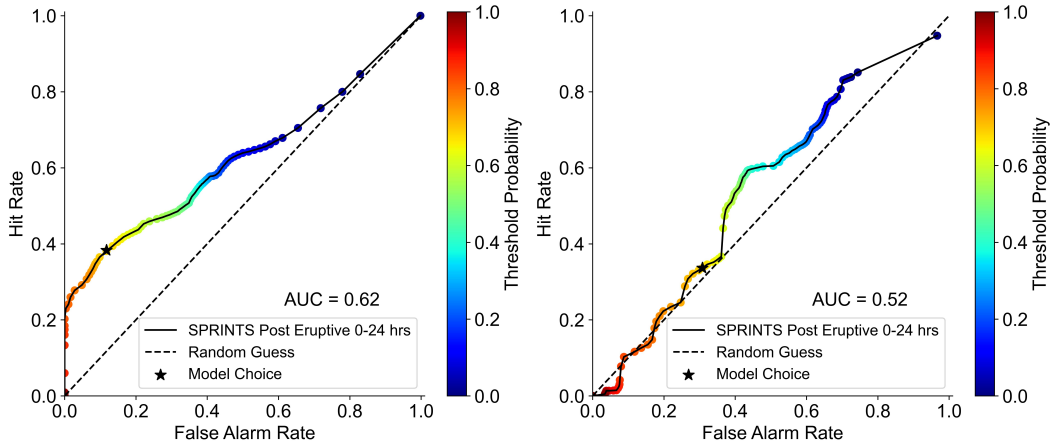
SPRINTS Probability	SEPVAL		Scoreboard	
	> 10 MeV ($N = 10,891$)	> 100 MeV ($N = 11,994$)	> 10 MeV ($N = 7192$)	> 100 MeV ($N = 7495$)
Brier Score	0.30	0.29	0.0058	0.0018
Brier Skill Score	0.34	−1.45	0.03	0.10
Brier (SEP)	0.45	0.39	0.78	0.70
Median P SEP	0.49	0.58	0.01	0.03
Area Under the Curve	0.62	0.55	0.76	0.90

Table 6.29: SPRINTS Probability Metrics.

is not necessarily an appropriate climatology for flare-triggered models, nonetheless, the scores are provided here for reference. SPRINTS shows low skill compared to this climatology for the >100 MeV channel for SEPVAL (BSS of −1.45), as well as for both energy channels on the SEP Scoreboards (BSS of 0.03 for >10 MeV and 0.10 for >100 MeV).

Figures 6.17 (SEPVAL) and 6.18 (Scoreboard) show the ROC curves for each dataset and energy channel. From these curves, the area underneath the model line is called the AUC, which can show model skill compared to a random guess, which would be correct 50% of the time. In evaluating this metric, the Scoreboards show much better ROC performance compared to SEPVAL, contrary to what was seen evaluating All Clear. Figure 6.19 shows the distribution of probabilities for Scoreboard forecasts at both energies for forecasts matching SEP events and those that do not. The figure shows median values shifting upwards only a few percentage points (0.01 and 0.03 for >10 and >100 MeV respectively). This means that Scoreboard SPRINTS does not strongly react to the changing SEP environment and predict an SEP event to occur. Improvements in the model are possible, however, the ROC curves in Figure 6.18 show that a higher Hit Rate is achievable for the >100 MeV channel, by changing the internal probability threshold to $\sim 50\%$. The same cannot be done to the same degree to the >10 MeV, where the max Hit Rate – while keeping False Alarm Rate low – is only $\sim 60\%$. The forecast probability distribution from SEPVAL shown in Figure 6.16 is quite distinct from the Scoreboard. In this case, quite high probabilities are shown for forecasts matching SEP events, with median values at 49% for > 10 MeV and 58% for > 100 MeV. The distribution for non-event periods is skewed towards significantly lower values: 12% for > 10 MeV and 36% for > 100 MeV. This resulted in a good hit rate and relatively low false alarm ratio for SEPVAL, much better than for the Scorebaord.

Our analysis shows that the SPRINTS model evaluated in SEPVAL has very different performance characteristics than the model evaluated on the Scoreboard. Part of this can be attributed to the different modes of operation, with SEPVAL SPRINTS producing forecasts every minute during the flare, while on the Scoreboard there is only one forecast per flare. The flare analysis is also different in the two environments, with the SEPVAL forecasts utilizing the model developer's own X-ray flux analyzer, while on the Scoreboard the final flare parameters from the LMSAL SolarSoft "Latest Events" web page are used (source: private communication with



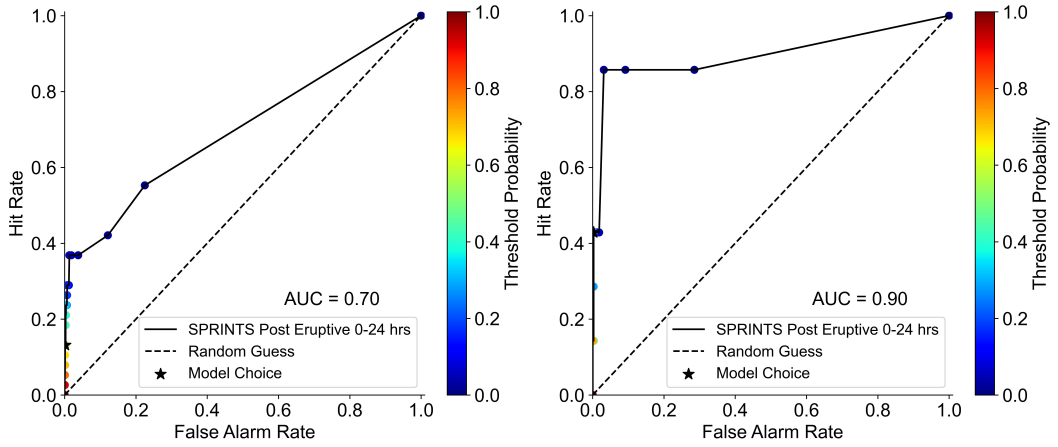
(a) $>$ 10 MeV (b) $>$ 100 MeV

Figure 6.17: SPRINTS ROC Curves for the SEPVAL Challenge.

model developers). Detailed analysis of the forecast JSONs for events covered in both SEPVAL and the Scoreboard revealed instances of different forecast probability for the same reported input data, indicating deeper differences in the models. Each of these factors leads us to conclude that the SEPVAL and Scoreboard versions of SPRINTS should be treated as different models and the results cannot be directly compared.

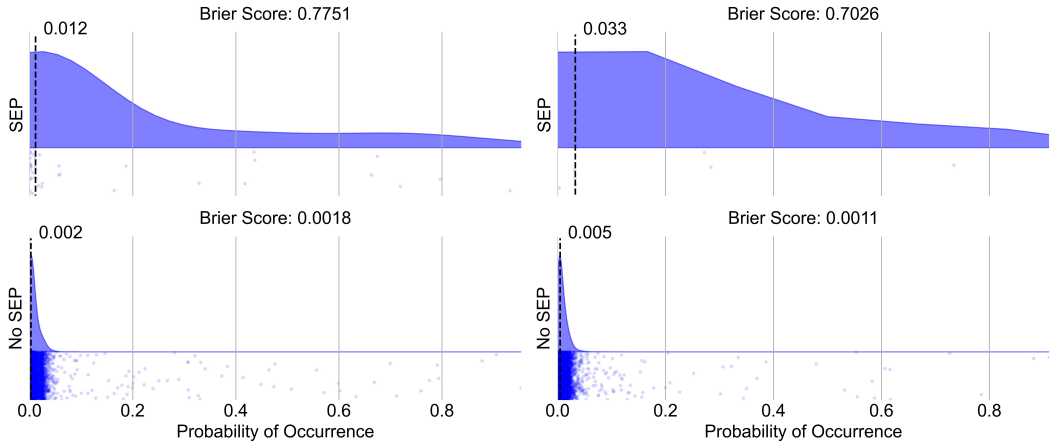
In summary:

- Evaluation of the SEPVAL and Scoreboard datasets show remarkably different performance. This together with known differences in model implementation between the two environments makes it impossible to compare performance derived from the two datasets.
- Performance on the SEP Scoreboards show that the model does not strongly react to the changing SEP environment and tends to issue low percent probabilities in advance of SEP events.
- The false alarm ratio on the Scoreboards is very high and hit rate is low, meaning the model is reacting to the wrong triggers. However the AUC is high, indicating that the model might be improved by lowering the probability threshold for forecasting Not Clear.



(a) $>$ 10 MeV (b) $>$ 100 MeV

Figure 6.18: SPRINTS ROC Curves for the SEP Scoreboards.



(a) >10 MeV (b) >100 MeV

Figure 6.19: Forecasted probabilities and Gaussian kernel density estimates for SPRINTS from the SEP Scoreboards, separated by event periods and non-event periods

6.6 SAWS-ASPECS

SAWS-ASPECS is a model with the aim to predict SEP event characteristics prior to precursors (i.e., flare, CME, rise in proton flux), then make updated predictions after the precursors using the newly available information (i.e., flare parameters, CME parameters, current proton flux). As such, the model is composed of two main pipelines: forecast and nowcast. Each pipeline is composed of a combination of modules "cfchars", "forsep", "prosper", "sepchars", and "sepprofiles" in addition to many helper modules.

The "cfchars" module uses SDO/Helioseismic and Magnetic Imager (HMI) magnetograms and calculates the effective connected magnetic field strength, B_{eff} (Georgoulis and Rust, 2007) for each active region. Using the value of B_{eff}, the module derives the 6, 12, 24, 48, and 72-hour flare cumulative probabilities for 28 GOES flare classes between C1.0 and X10.0+. This calculation uses fitting similar to those from the FORSPEF database (Anastasiadis et al., 2017; Papaioannou et al., 2015; Papaioannou, Athanasios et al., 2016) for a 24-hour window using historic flares in each of the flare bins. For each of these flare classes and forecast windows, the CME likelihood is calculated. Additionally, the module predicts the CME speed from B_{eff} (Georgoulis, Manolis K. et al., 2021).

The "forsep" module uses the longitude of each active region derived from SDO/HMI magnetograms in the "cfchars" module. Using the FORSPEF database, it gets the closest 18 historic flares west and 18 historic flares east of each region. The fraction of flares resulting in SEP events is calculated for 6 flare magnitude bins (C1-C4, C4-C7, C7-M1, M1-M6, M6-X1, and X10+). The flaring probability for each region resulting from the "cfchars" module is rebinned into these 6 flare bins. Cumulative distribution functions are calculated for each bin used to calculate the probability of an SEP event occurring after summing across all flare bins.

The "prosper" module calculates the probability of SEP event occurrence using the results of Papaioannou et al. (2022). The paper derived the cumulative distribution function for flares in two longitudinal bins ($<$20 deg and $\geq$20 deg), for CMEs in three bins of CME width (0-119 deg, 120-359 deg, and 360 deg), and for both flares and CMEs together using the binning from each. Probability distribution functions are calculated from the cumulative distribution functions. Historic events from the FORSPEF database are used to calculate the fraction of flares or CMEs in each bin that resulted in an SEP event and those that resulted in no SEP event. Then a Bayesian technique is used to calculate the conditional probability of SEP occurrence given the flare or CME parameters.

The "sepchars" module makes predictions of the peak SEP flux also based on a method derived in Papaioannou et al. (2022). The method derived cumulative distribution functions of peak proton flux for $\geq$10 MeV, $\geq$30 MeV, $\geq$100 MeV, $\geq$300 MeV. The fits were derived for different conditions based on each energy reaching a certain threshold and for different flare and CME parameters. Within the module, the peak flux is calculated using a weighted average between the probability of flux reaching a certain threshold derived from the "forsep" or "prosper" module and the background proton intensity derived from the FORSPEF database, using

the probability of SEP occurrence as the weight. This is done for the 50% and 90% confidence levels, interpreted as 50% and 90% confidence that the observed peak flux will be equal to or less than the predicted value.

The "sepprofiles" module predicts a proton intensity time profile using one of two methods: Kahler-Ling (Kahler and Ling, 2018) or SOLPENCO2 (Aran et al., 2006). For the Kahler-Ling method, a modified Weibull function is fit to the real-time GOES proton intensity. If there is insufficient proton data, then the module falls back to the default SOLPENCO2 method. The SOLPENCO2 method includes previously-ran proton intensity time profiles using the SOLPENCO2 model for a variety of source conditions. In this case, the flare parameters are used to select the best fitting SOLPENCO2 time profile. For both methods, the best time profile is chosen based on the least chi-squared value, then scaled to the predicted peak.

The forecast pipeline runs every 3 hours. It starts with the "cfchars" and "forsep" modules to produce the probability of an SEP event. These predictions are reflected on the SEP Scoreboard as the "SAWS-ASPECS 0-X hrs" variant, where X here is the forecast window. This probability is then fed into the forecast pipeline version of the "sepchars" and "sepprofiles" modules to predict a peak flux and time profile. On the SEP Scoreboards, these variants are "SAWS-ASPECS 0-X hrs 50%/90%".

The nowcast pipeline runs every 5 minutes and first checks the database for recent flares and CMEs. If none occurred, then a zero probability is produced. On the SEP Scoreboard, this prediction is the "SAWS-ASPECS" variant and commonly referred to as "untriggered nowcast". If the pipeline instead finds recent flares or CMEs, these parameters are fed into the "prosper" module to produce the probability of SEP occurrence. These predictions show up on the SEP Scoreboard as "SAWS-ASPECS Flare", "SAWS-ASPECS CME", and "SAWS-ASPECS Flare+CME". The probabilities from the "prosper" module are then fed into the "sepchars" module to produce a prediction for peak flux, then the "sepprofiles" module to predict a time profile based on the predicted peak flux. If the predictions are only based on CME parameters, then only a constant profile is predicted. The time profile predictions show up on the SEP Scoreboard as "SAWS-ASPECS Flare 50%/90%", "SAWS-ASPECS CME 50%/90%", and "SAWS-ASPECS Flare+CME 50%/90%".

One of the helper modules that runs at the end of either pipeline helps derive the All Clear predictions based on the probability predictions and time-profile predictions. For the probability predictions, a threshold is applied depending on the energy. For $\geq$10 MeV, $\geq$30 MeV, $\geq$100 MeV, and $\geq$300 MeV predictions, the thresholds are 26%, 20%, 15%, and 12%, respectively. These thresholds were previously derived from optimizing skill scores. For the time-profile predictions, All Clear is derived from whether the peak of the time-profile is above or below a threshold depending on the energy. For $\geq$10 MeV, $\geq$30 MeV, $\geq$100 MeV, and $\geq$300 MeV predictions, the thresholds are 10 pfu, 3 pfu, 1 pfu, and 0.3 pfu, respectively. These thresholds are what are used in operations and were provided to the modelers.

For the SEPVAL results, predictions using the "electrons" module is also used within the nowcast pipeline. This module reads in real-time electron data from Advanced Composition Explorer (ACE)/Electron, Proton, and Alpha Monitor (EPAM)

and uses convolution to identify abrupt changes in the time series. If this criteria for an electron event is not met, then the probability of SEP occurrence produced by the "prosper" module is set to zero. This module is not included in the version running on the SEP Scoreboard.

ASPECS is not a pure prediction model. The output is a combination of predictions and observations. The goal of ASPECS is to answer the question, "what is the biggest threat right now?" As such, if the predicted peak flux is lower than the current flux, then ASPECS outputs the current flux. This is not a true prediction and makes validating the model very challenging – specifically for the SEP Scoreboard where validation relies on automated processes. Therefore, the SEP Scoreboard results are not included for ASPECS. Since the SEPVAL results were carefully considered, they are discussed below. The current version of ASPECS on the SEP Scoreboard (as of May 2025) will be updated with a version that excludes this interplay between predictions and observations, and therefore, this new version will be evaluated in a future study once the SEP Scoreboard collects enough predictions.

Table 6.30 shows the All Clear metrics for ≥10 MeV predictions for the SEPVAL results, and Table 6.31 shows the same but for variants of SAWS-ASPECS that include electron flux. The SEPVAL challenge provided four different sets of CME parameters for model developers to use as appropriate for their model. The SAWS-ASPECS group used all sets as an interesting exercise. In the tables, SOHO inputs are hand-derived 2D plane-of-sky measurements provided in the CDAW LASCO CME catalog. M2M are 3D CME parameters measured by M2M analysts for the SEPVAL challenge and listed in DONKI. CACTus are automated 2D plane-of-sky measurements where CACTus1 is the median speed and CACTus2 is the maximum CME speed. Predictions using median CME speeds from CACTus (CACTus1) under-perform compared to using maximum CME speeds (CACTus2). Predictions improve when using CMEs from the M2M catalog or from SOHO. Using SOHO CMEs gives a high Hit Rate but also a high False Alarm Rate, whereas using M2M CMEs gives a lower Hit Rate but also a much lower False Alarm Rate. Including flare information with SOHO CMEs increases the False Alarm Rate even further for probability predictions. In general, variants only using CME information are the least performant, using only flare information performs well, and using both flare and CME information performs similarly or better than only using flare information. Including electron flux drastically reduces False Alarm Rates, with some rates dropping to zero and others being reduced by up to 50%. The True Skill Statistics consequentially nearly doubles. These results compare well with the independent validation study using the SEPVAL event list (Papaioannou et al., 2025).

The probabilistic metrics of the SEPVAL results are in Tables 6.34 and 6.35. The results show a clear advantage to using SOHO CMEs as input rather than CACTus when distinguishing between SEP events and non-SEP events. This is evident in the 0.17-028 increase in AUC value for ≥10 MeV and seen in the ROC plot (Figure 6.20). The AUC values when using CACTus CMEs are improved by 0.15-0.18 when including recent flare parameters. However, flare parameters do not improve predictions when using SOHO CMEs as input. Including electrons to reduce

the number of false alarms improves AUC values for every variant of SAWS-ASPECS by 0.07-0.33.

Tables 6.36 and 6.37 show onset peak flux metrics for bias, accuracy, and correlation for the SEPVAL results. It should be noted that the incorporation of electrons does not affect these metrics much since the module only nullifies predictions in the case where electron flux does not rise and therefore no protons follow, whereas the onset peak metrics can only be calculated when an SEP event occurs. The bias is also represented graphically in Figures 6.22 and 6.23. For $\geq$10 MeV predictions, the pure CME variants of SAWS-ASPECS under-predict on average by about two orders of magnitude and perform similarly compared to each other. When including flare parameters, most predictions are within one order of magnitude, with the 90% confidence level variants predicting higher than the 50% confidence level variants. Figures 6.24 and 6.25 show the cumulative distribution functions for absolute log error. For the SEPVAL results, the flare+CME (M2M) variant performs the best with about 67% of the predictions being within one order of magnitude error. This is true for the 50% confidence level predictions. For the 90% confidence level predictions, only about 25% are within one order of magnitude error. These 90% confidence level variants tend to over-predict, as seen by the positive median log error values in Table 6.36. This is by design since the model was trained on the larger SEP events of solar cycle 23. About 44% of the 50% confidence level predictions for the flare-only variant fall within one order of magnitude error. For the pure CME variants, no predictions are within one order of magnitude error. The pure CME variants are also least correlated, with SCC values ranging from 0.18-0.23. The flare-only variants also fall into this range. Predictions including combined flare and CME information are slightly more correlated, with SSC values ranging from 0.27-0.60, where the M2M variant performing the best.

Model	TP	FP	TN	FN	Total	PC	H	F	FAR	TS	TSS	HSS
SAWS-ASPECS CME (CACTUS1)	2	2	25	27	56	0.48	0.07	0.07	0.50	0.06	-0.01	-0.00
SAWS-ASPECS CME (CACTUS1) 50%	1	0	27	28	56	0.50	0.03	0.00	0.00	0.03	0.03	0.03
SAWS-ASPECS CME (CACTUS1) 90%	1	0	27	28	56	0.50	0.03	0.00	0.00	0.03	0.03	0.03
SAWS-ASPECS CME (CACTUS2)	11	6	21	18	56	0.57	0.38	0.22	0.35	0.31	0.16	0.15
SAWS-ASPECS CME (CACTUS2) 50%	1	0	27	28	56	0.50	0.03	0.00	0.00	0.03	0.03	0.03
SAWS-ASPECS CME (CACTUS2) 90%	1	0	27	28	56	0.50	0.03	0.00	0.00	0.03	0.03	0.03
SAWS-ASPECS CME (M2M)	6	1	29	21	57	0.61	0.22	0.03	0.14	0.21	0.19	0.20
SAWS-ASPECS CME (M2M) 50%	0	0	30	27	57	0.53	0.00	0.00	nan	0.00	0.00	0.00
SAWS-ASPECS CME (M2M) 90%	0	0	30	27	57	0.53	0.00	0.00	nan	0.00	0.00	0.00
SAWS-ASPECS CME (SOHO)	29	13	17	3	62	0.74	0.91	0.43	0.31	0.64	0.47	0.48
SAWS-ASPECS CME (SOHO) 50%	1	0	30	31	62	0.50	0.03	0.00	0.00	0.03	0.03	0.03
SAWS-ASPECS CME (SOHO) 90%	1	0	30	31	62	0.50	0.03	0.00	0.00	0.03	0.03	0.03
SAWS-ASPECS flare	21	12	18	8	59	0.66	0.72	0.40	0.36	0.51	0.32	0.32
SAWS-ASPECS flare 50%	21	12	18	8	59	0.66	0.72	0.40	0.36	0.51	0.32	0.32
SAWS-ASPECS flare 90%	21	12	18	8	59	0.66	0.72	0.40	0.36	0.51	0.32	0.32
SAWS-ASPECS flare + CME (CACTUS1)	20	9	18	6	53	0.72	0.77	0.33	0.31	0.57	0.44	0.43
SAWS-ASPECS flare + CME (CACTUS1) 50%	14	6	21	12	53	0.66	0.54	0.22	0.30	0.44	0.32	0.32
SAWS-ASPECS flare + CME (CACTUS1) 90%	17	7	20	9	53	0.70	0.65	0.26	0.29	0.52	0.39	0.40
SAWS-ASPECS flare + CME (CACTUS2)	23	11	16	3	53	0.74	0.88	0.41	0.32	0.62	0.48	0.47
SAWS-ASPECS flare + CME (CACTUS2) 50%	21	7	20	5	53	0.77	0.81	0.26	0.25	0.64	0.55	0.55
SAWS-ASPECS flare + CME (CACTUS2) 90%	22	10	17	4	53	0.74	0.85	0.37	0.31	0.61	0.48	0.47
SAWS-ASPECS flare + CME (M2M)	18	11	19	6	54	0.69	0.75	0.37	0.38	0.51	0.38	0.38
SAWS-ASPECS flare + CME (M2M) 50%	17	7	23	7	54	0.74	0.71	0.23	0.29	0.55	0.47	0.47
SAWS-ASPECS flare + CME (M2M) 90%	17	10	20	7	54	0.69	0.71	0.33	0.37	0.50	0.38	0.37
SAWS-ASPECS flare + CME (SOHO)	26	21	9	3	59	0.59	0.90	0.70	0.45	0.52	0.20	0.19
SAWS-ASPECS flare + CME (SOHO) 50%	25	16	14	4	59	0.66	0.86	0.53	0.39	0.56	0.33	0.33
SAWS-ASPECS flare + CME (SOHO) 90%	26	18	12	3	59	0.64	0.90	0.60	0.41	0.55	0.30	0.29

Table 6.30: All Clear metrics for SAWS-ASPECS $\geq$10 MeV SEPVAL predictions. The metrics include True Positives (TP), False Positives (FP), True Negatives (TN), False Negatives (FN), Total, Percent Correct (PC), Hit Rate (H), False Alarm Rate (F), False Alarm Ratio (FAR), Threat Score (TS), True Skill Statistic (TSS), and Heidke Skill Score (HSS).

SAWS-ASPECS CME (CACTUS1) electrons	2	0	27	26	55	0.53	0.07	0.00	0.00	0.07	0.07	0.07
SAWS-ASPECS CME (CACTUS1) electrons 50%	0	0	27	27	54	0.50	0.00	0.00	nan	0.00	0.00	0.00
SAWS-ASPECS CME (CACTUS1) electrons 90%	0	0	27	27	54	0.50	0.00	0.00	nan	0.00	0.00	0.00
SAWS-ASPECS CME (CACTUS2) electrons	10	1	26	18	55	0.65	0.36	0.04	0.09	0.34	0.32	0.32
SAWS-ASPECS CME (CACTUS2) electrons 50%	0	0	27	27	54	0.50	0.00	0.00	nan	0.00	0.00	0.00
SAWS-ASPECS CME (CACTUS2) electrons 90%	0	0	27	27	54	0.50	0.00	0.00	nan	0.00	0.00	0.00
SAWS-ASPECS CME (M2M) electrons	6	0	30	21	57	0.63	0.22	0.00	0.00	0.22	0.22	0.23
SAWS-ASPECS CME (M2M) electrons 50%	0	0	30	26	56	0.54	0.00	0.00	nan	0.00	0.00	0.00
SAWS-ASPECS CME (M2M) electrons 90%	0	0	30	26	56	0.54	0.00	0.00	nan	0.00	0.00	0.00
SAWS-ASPECS CME (SOHO) electrons	27	2	28	4	61	0.90	0.87	0.07	0.07	0.82	0.80	0.80
SAWS-ASPECS CME (SOHO) electrons 50%	0	0	30	30	60	0.50	0.00	0.00	nan	0.00	0.00	0.00
SAWS-ASPECS CME (SOHO) electrons 90%	0	0	30	30	60	0.50	0.00	0.00	nan	0.00	0.00	0.00
SAWS-ASPECS flare + CME (CACTUS1) electrons	18	3	24	7	52	0.81	0.72	0.11	0.14	0.64	0.61	0.61
SAWS-ASPECS flare + CME (CACTUS1) electrons 50%	13	2	25	11	51	0.75	0.54	0.07	0.13	0.50	0.47	0.48
SAWS-ASPECS flare + CME (CACTUS1) electrons 90%	15	2	25	9	51	0.78	0.62	0.07	0.12	0.58	0.55	0.56
SAWS-ASPECS flare + CME (CACTUS2) electrons	21	3	24	4	52	0.87	0.84	0.11	0.12	0.75	0.73	0.73
SAWS-ASPECS flare + CME (CACTUS2) electrons 50%	20	3	24	4	51	0.86	0.83	0.11	0.13	0.74	0.72	0.72
SAWS-ASPECS flare + CME (CACTUS2) electrons 90%	20	3	24	4	51	0.86	0.83	0.11	0.13	0.74	0.72	0.72
SAWS-ASPECS flare + CME (M2M) electrons	18	3	27	6	54	0.83	0.75	0.10	0.14	0.67	0.65	0.66
SAWS-ASPECS flare + CME (M2M) electrons 50%	17	2	28	6	53	0.85	0.74	0.07	0.11	0.68	0.67	0.69
SAWS-ASPECS flare + CME (M2M) electrons 90%	17	2	28	6	53	0.85	0.74	0.07	0.11	0.68	0.67	0.69
SAWS-ASPECS flare + CME (SOHO) electrons	24	6	24	4	58	0.83	0.86	0.20	0.20	0.71	0.66	0.66
SAWS-ASPECS flare + CME (SOHO) electrons 50%	23	5	25	4	57	0.84	0.85	0.17	0.18	0.72	0.69	0.68
SAWS-ASPECS flare + CME (SOHO) electrons 90%	24	6	24	3	57	0.84	0.89	0.20	0.20	0.73	0.69	0.69
SAWS-ASPECS flare electrons	19	3	27	9	58	0.79	0.68	0.10	0.14	0.61	0.58	0.58
SAWS-ASPECS flare electrons 50%	18	3	27	9	57	0.79	0.67	0.10	0.14	0.60	0.57	0.57
SAWS-ASPECS flare electrons 90%	19	3	27	8	57	0.81	0.70	0.10	0.14	0.63	0.60	0.61

Table 6.31: All Clear metrics for SAWS-ASPECS ≥10 MeV SEPVAL predictions (electron models). The metrics include True Positives (TP), False Positives (FP), True Negatives (TN), False Negatives (FN), Total, Percent Correct (PC), Hit Rate (H), False Alarm Rate (F), False Alarm Ratio (FAR), Threat Score (TS), True Skill Statistic (TSS), and Heidke Skill Score (HSS).

Model	TP	FP	TN	FN	Total	PC	H	F	FAR	TS	TSS	HSS
SAWS-ASPECS CME (CACTUS1)	2	1	43	11	57	0.79	0.15	0.02	0.33	0.14	0.13	0.18
SAWS-ASPECS CME (CACTUS1) 50%	0	0	44	13	57	0.77	0.00	0.00	nan	0.00	0.00	0.00
SAWS-ASPECS CME (CACTUS1) 90%	0	0	44	13	57	0.77	0.00	0.00	nan	0.00	0.00	0.00
SAWS-ASPECS CME (CACTUS2)	4	9	35	9	57	0.68	0.31	0.20	0.69	0.18	0.10	0.10
SAWS-ASPECS CME (CACTUS2) 50%	0	0	44	13	57	0.77	0.00	0.00	nan	0.00	0.00	0.00
SAWS-ASPECS CME (CACTUS2) 90%	0	0	44	13	57	0.77	0.00	0.00	nan	0.00	0.00	0.00
SAWS-ASPECS CME (M2M)	3	0	45	10	58	0.83	0.23	0.00	0.00	0.23	0.23	0.32
SAWS-ASPECS CME (M2M) 50%	0	0	45	13	58	0.78	0.00	0.00	nan	0.00	0.00	0.00
SAWS-ASPECS CME (M2M) 90%	0	0	45	13	58	0.78	0.00	0.00	nan	0.00	0.00	0.00
SAWS-ASPECS CME (SOHO)	13	19	27	4	63	0.63	0.76	0.41	0.59	0.36	0.35	0.28
SAWS-ASPECS CME (SOHO) 50%	0	0	46	17	63	0.73	0.00	0.00	nan	0.00	0.00	0.00
SAWS-ASPECS CME (SOHO) 90%	0	0	46	17	63	0.73	0.00	0.00	nan	0.00	0.00	0.00
SAWS-ASPECS flare	13	18	27	2	60	0.67	0.87	0.40	0.58	0.39	0.47	0.34
SAWS-ASPECS flare 50%	3	4	41	12	60	0.73	0.20	0.09	0.57	0.16	0.11	0.14
SAWS-ASPECS flare 90%	10	15	30	5	60	0.67	0.67	0.33	0.60	0.33	0.33	0.27
SAWS-ASPECS flare + CME (CACTUS1)	5	14	29	6	54	0.63	0.45	0.33	0.74	0.20	0.13	0.10
SAWS-ASPECS flare + CME (CACTUS1) 50%	2	4	39	9	54	0.76	0.18	0.09	0.67	0.13	0.09	0.11
SAWS-ASPECS flare + CME (CACTUS1) 90%	4	8	35	7	54	0.72	0.36	0.19	0.67	0.21	0.18	0.17
SAWS-ASPECS flare + CME (CACTUS2)	6	16	27	5	54	0.61	0.55	0.37	0.73	0.22	0.17	0.13
SAWS-ASPECS flare + CME (CACTUS2) 50%	5	4	39	6	54	0.81	0.45	0.09	0.44	0.33	0.36	0.39
SAWS-ASPECS flare + CME (CACTUS2) 90%	5	14	29	6	54	0.63	0.45	0.33	0.74	0.20	0.13	0.10
SAWS-ASPECS flare + CME (M2M)	8	10	34	3	55	0.76	0.73	0.23	0.56	0.38	0.50	0.40
SAWS-ASPECS flare + CME (M2M) 50%	2	1	43	9	55	0.82	0.18	0.02	0.33	0.17	0.16	0.22
SAWS-ASPECS flare + CME (M2M) 90%	7	7	37	4	55	0.80	0.64	0.16	0.50	0.39	0.48	0.43
SAWS-ASPECS flare + CME (SOHO)	14	25	20	1	60	0.57	0.93	0.56	0.64	0.35	0.38	0.25
SAWS-ASPECS flare + CME (SOHO) 50%	12	21	24	3	60	0.60	0.80	0.47	0.64	0.33	0.33	0.24
SAWS-ASPECS flare + CME (SOHO) 90%	13	22	23	2	60	0.60	0.87	0.49	0.63	0.35	0.38	0.26

Table 6.32: All Clear metrics for SAWS-ASPECS $\geq$100 MeV SEPVAL predictions. The metrics include True Positives (TP), False Positives (FP), True Negatives (TN), False Negatives (FN), Total, Percent Correct (PC), Hit Rate (H), False Alarm Rate (F), False Alarm Ratio (FAR), Threat Score (TS), True Skill Statistic (TSS), and Heidke Skill Score (HSS).

Model	TP	FP	TN	FN	Total	PC	H	F	FAR	TS	TSS	HSS
SAWS-ASPECS CME (CACTUS1) electrons	2	0	44	10	56	0.82	0.17	0.00	0.00	0.17	0.17	0.24
SAWS-ASPECS CME (CACTUS1) electrons 50%	0	0	44	12	56	0.79	0.00	0.00	nan	0.00	0.00	0.00
SAWS-ASPECS CME (CACTUS1) electrons 90%	0	0	44	12	56	0.79	0.00	0.00	nan	0.00	0.00	0.00
SAWS-ASPECS CME (CACTUS2) electrons	4	6	38	8	56	0.75	0.33	0.14	0.60	0.22	0.20	0.21
SAWS-ASPECS CME (CACTUS2) electrons 50%	0	0	44	12	56	0.79	0.00	0.00	nan	0.00	0.00	0.00
SAWS-ASPECS CME (CACTUS2) electrons 90%	0	0	44	12	56	0.79	0.00	0.00	nan	0.00	0.00	0.00
SAWS-ASPECS CME (M2M) electrons	3	0	45	9	57	0.84	0.25	0.00	0.00	0.25	0.25	0.34
SAWS-ASPECS CME (M2M) electrons 50%	0	0	45	12	57	0.79	0.00	0.00	nan	0.00	0.00	0.00
SAWS-ASPECS CME (M2M) electrons 90%	0	0	45	12	57	0.79	0.00	0.00	nan	0.00	0.00	0.00
SAWS-ASPECS CME (SOHO) electrons	12	14	32	4	62	0.71	0.75	0.30	0.54	0.40	0.45	0.37
SAWS-ASPECS CME (SOHO) electrons 50%	0	0	46	16	62	0.74	0.00	0.00	nan	0.00	0.00	0.00
SAWS-ASPECS CME (SOHO) electrons 90%	0	0	46	16	62	0.74	0.00	0.00	nan	0.00	0.00	0.00
SAWS-ASPECS flare + CME (CACTUS1) electrons	4	11	32	6	53	0.68	0.40	0.26	0.73	0.19	0.14	0.12
SAWS-ASPECS flare + CME (CACTUS1) electrons 50%	2	2	41	8	53	0.81	0.20	0.05	0.50	0.17	0.15	0.20
SAWS-ASPECS flare + CME (CACTUS1) electrons 90%	3	6	37	7	53	0.75	0.30	0.14	0.67	0.19	0.16	0.17
SAWS-ASPECS flare + CME (CACTUS2) electrons	5	13	30	5	53	0.66	0.50	0.30	0.72	0.22	0.20	0.15
SAWS-ASPECS flare + CME (CACTUS2) electrons 50%	4	2	41	6	53	0.85	0.40	0.05	0.33	0.33	0.35	0.42
SAWS-ASPECS flare + CME (CACTUS2) electrons 90%	4	11	32	6	53	0.68	0.40	0.26	0.73	0.19	0.14	0.12
SAWS-ASPECS flare + CME (M2M) electrons	7	10	34	3	54	0.76	0.70	0.23	0.59	0.35	0.47	0.37
SAWS-ASPECS flare + CME (M2M) electrons 50%	1	1	43	9	54	0.81	0.10	0.02	0.50	0.09	0.08	0.11
SAWS-ASPECS flare + CME (M2M) electrons 90%	5	6	38	5	54	0.80	0.50	0.14	0.55	0.31	0.36	0.35
SAWS-ASPECS flare + CME (SOHO) electrons	12	15	30	2	59	0.71	0.86	0.33	0.56	0.41	0.52	0.40
SAWS-ASPECS flare + CME (SOHO) electrons 50%	11	11	34	3	59	0.76	0.79	0.24	0.50	0.44	0.54	0.45
SAWS-ASPECS flare + CME (SOHO) electrons 90%	12	14	31	2	59	0.73	0.86	0.31	0.54	0.43	0.55	0.42
SAWS-ASPECS flare electrons	11	11	34	3	59	0.76	0.79	0.24	0.50	0.44	0.54	0.45
SAWS-ASPECS flare electrons 50%	2	4	41	12	59	0.73	0.14	0.09	0.67	0.11	0.05	0.07
SAWS-ASPECS flare electrons 90%	8	9	36	6	59	0.75	0.57	0.20	0.53	0.35	0.37	0.35

Table 6.33: All Clear metrics for SAWS-ASPECS $\geq$100 MeV SEPVAL predictions (electron models). The metrics include True Positives (TP), False Positives (FP), True Negatives (TN), False Negatives (FN), Total, Percent Correct (PC), Hit Rate (H), False Alarm Rate (F), False Alarm Ratio (FAR), Threat Score (TS), True Skill Statistic (TSS), and Heidke Skill Score (HSS).

Model	BS	BSS	AUC
SAWS-ASPECS CME (CACTUS1)	0.45	0.08	0.56
SAWS-ASPECS CME (CACTUS2)	0.36	0.26	0.60
SAWS-ASPECS CME (M2M)	0.36	0.20	0.77
SAWS-ASPECS CME (SOHO)	0.18	0.63	0.84
SAWS-ASPECS flare	0.27	0.41	0.69
SAWS-ASPECS flare + CME (CACTUS1)	0.26	0.43	0.74
SAWS-ASPECS flare + CME (CACTUS2)	0.24	0.49	0.75
SAWS-ASPECS flare + CME (M2M)	0.19	0.53	0.80
SAWS-ASPECS flare + CME (SOHO)	0.28	0.40	0.83
SAWS-ASPECS flare + CME (CACTUS1) electrons	0.19	0.57	0.89
SAWS-ASPECS flare + CME (CACTUS2) electrons	0.15	0.67	0.90
SAWS-ASPECS flare + CME (M2M) electrons	0.14	0.65	0.94
SAWS-ASPECS flare + CME (SOHO) electrons	0.11	0.75	0.93
SAWS-ASPECS flare electrons	0.20	0.56	0.90

Table 6.34: Probabilistic metrics for SAWS-ASPECS ≥10 MeV SEPVAL predictions. The metrics include Brier Score (BS), Brier Skill Score (BSS), and Area Under the ROC (AUC).

Model	BS	BSS	AUC
SAWS-ASPECS CME (CACTUS1)	0.21	0.00	0.52
SAWS-ASPECS CME (CACTUS2)	0.20	0.05	0.53
SAWS-ASPECS CME (M2M)	0.20	0.07	0.70
SAWS-ASPECS CME (SOHO)	0.17	0.32	0.76
SAWS-ASPECS flare	0.17	0.28	0.76
SAWS-ASPECS flare + CME (CACTUS1)	0.20	-0.03	0.54
SAWS-ASPECS flare + CME (CACTUS2)	0.19	0.03	0.59
SAWS-ASPECS flare + CME (M2M)	0.13	0.33	0.78
SAWS-ASPECS flare + CME (SOHO)	0.15	0.35	0.82
SAWS-ASPECS flare + CME (CACTUS1) electrons	0.19	-0.07	0.57
SAWS-ASPECS flare + CME (CACTUS2) electrons	0.17	0.04	0.61
SAWS-ASPECS flare + CME (M2M) electrons	0.13	0.27	0.79
SAWS-ASPECS flare + CME (SOHO) electrons	0.15	0.34	0.82
SAWS-ASPECS flare electrons	0.17	0.24	0.79

Table 6.35: Probabilistic metrics for SAWS-ASPECS ≥100 MeV SEPVAL predictions. The metrics include Brier Score (BS), Brier Skill Score (BSS), and Area Under the ROC (AUC).

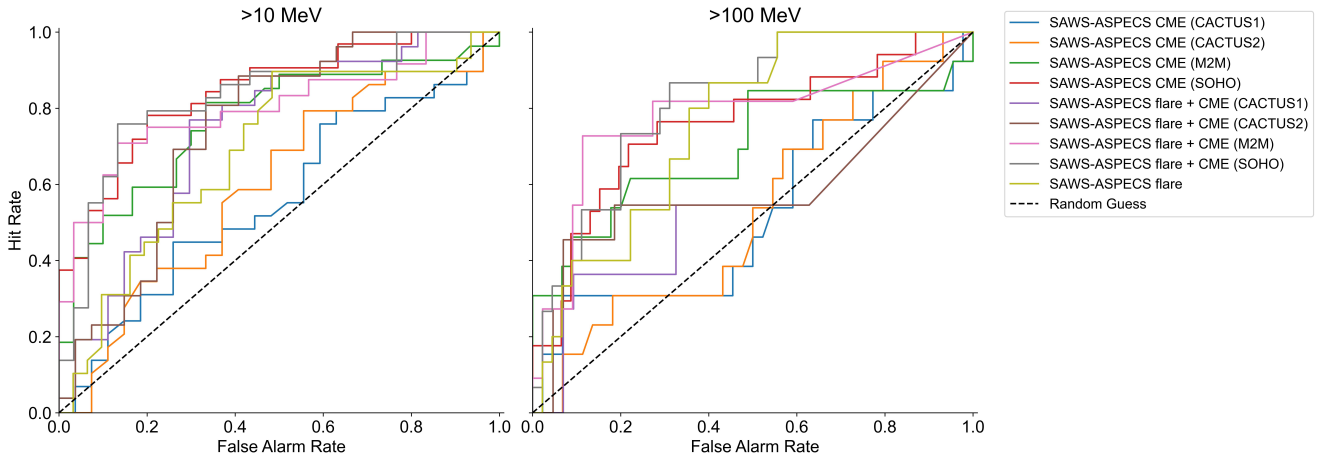


Figure 6.20: ROC curve for SAWS-ASPECS ≥10 MeV and ≥100 MeV SEPVAL predictions. Plot constructed with 10 linear-spaced bins.

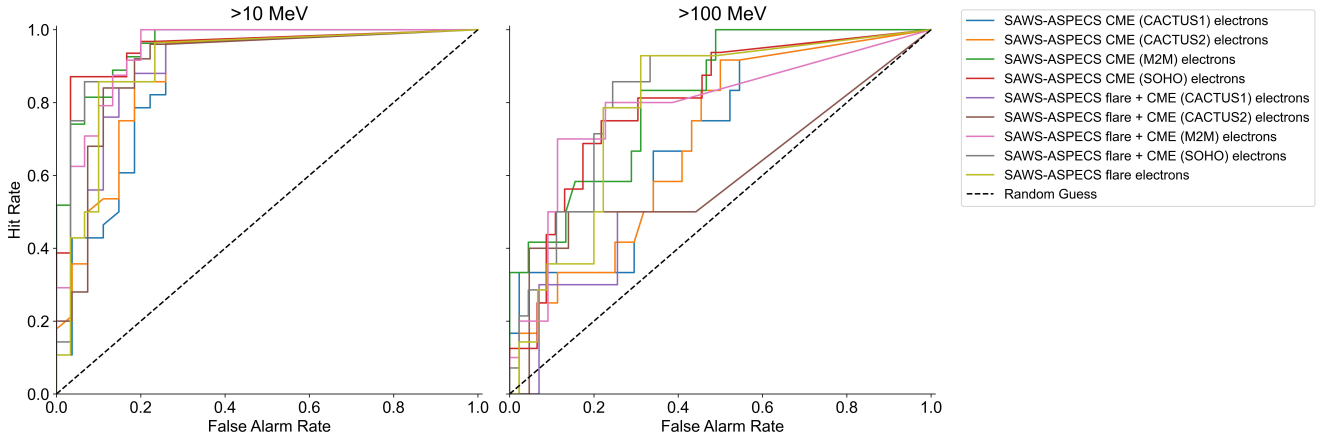


Figure 6.21: ROC curve for SAWS-ASPECS ≥10 MeV and ≥100 MeV SEPVAL predictions for variants including electron flux. Plot constructed with 10 linear-spaced bins.

Model	MedLE	MedALE	SCC
SAWS-ASPECS CME (CACTUS1) 50%	-2.06	2.06	0.19
SAWS-ASPECS CME (CACTUS1) 90%	-1.99	1.99	0.20
SAWS-ASPECS CME (CACTUS2) 50%	-2.06	2.06	0.19
SAWS-ASPECS CME (CACTUS2) 90%	-1.99	1.99	0.20
SAWS-ASPECS CME (M2M) 50%	-2.01	2.01	0.18
SAWS-ASPECS CME (M2M) 90%	-1.97	1.97	0.19
SAWS-ASPECS CME (SOHO) 50%	-2.10	2.10	0.21
SAWS-ASPECS CME (SOHO) 90%	-2.03	2.03	0.23
SAWS-ASPECS flare 50%	-0.67	1.25	0.21
SAWS-ASPECS flare 90%	0.96	1.62	0.21
SAWS-ASPECS flare + CME (CACTUS1) 50%	-1.20	1.29	0.27
SAWS-ASPECS flare + CME (CACTUS1) 90%	-0.21	1.08	0.30
SAWS-ASPECS flare + CME (CACTUS2) 50%	-0.21	0.89	0.30
SAWS-ASPECS flare + CME (CACTUS2) 90%	1.06	1.44	0.28
SAWS-ASPECS flare + CME (M2M) 50%	-0.08	0.49	0.60
SAWS-ASPECS flare + CME (M2M) 90%	1.07	1.41	0.59
SAWS-ASPECS flare + CME (SOHO) 50%	0.76	0.99	0.35
SAWS-ASPECS flare + CME (SOHO) 90%	2.11	2.24	0.37
SAWS-ASPECS flare + CME (CACTUS1) electrons 50%	-0.89	1.16	0.27
SAWS-ASPECS flare + CME (CACTUS1) electrons 90%	-0.20	1.03	0.30
SAWS-ASPECS flare + CME (CACTUS2) electrons 50%	-0.21	0.77	0.26
SAWS-ASPECS flare + CME (CACTUS2) electrons 90%	1.06	1.48	0.23
SAWS-ASPECS flare + CME (M2M) electrons 50%	-0.08	0.41	0.63
SAWS-ASPECS flare + CME (M2M) electrons 90%	1.06	1.31	0.63
SAWS-ASPECS flare + CME (SOHO) electrons 50%	0.74	0.89	0.33
SAWS-ASPECS flare + CME (SOHO) electrons 90%	2.09	2.11	0.34
SAWS-ASPECS flare electrons 50%	-0.81	1.23	0.16
SAWS-ASPECS flare electrons 90%	0.80	1.75	0.17

Table 6.36: Onset peak metrics for SAWS-ASPECS $\geq$10 MeV SEPVAL predictions. The metrics include Median Log Error (MedLE), Median Absolute Log Error (MedALE), and Spearman Correlation Coefficient (SCC).

Model	MedLE	MedALE	SCC
SAWS-ASPECS CME (CACTUS1) 50%	-1.74	1.74	-0.31
SAWS-ASPECS CME (CACTUS1) 90%	-1.63	1.63	-0.28
SAWS-ASPECS CME (CACTUS2) 50%	-1.74	1.74	-0.31
SAWS-ASPECS CME (CACTUS2) 90%	-1.63	1.63	-0.28
SAWS-ASPECS CME (M2M) 50%	-1.86	1.86	-0.21
SAWS-ASPECS CME (M2M) 90%	-1.77	1.77	-0.31
SAWS-ASPECS CME (SOHO) 50%	-1.86	1.86	-0.11
SAWS-ASPECS CME (SOHO) 90%	-1.77	1.77	-0.30
SAWS-ASPECS flare 50%	-0.79	0.99	0.02
SAWS-ASPECS flare 90%	0.10	0.60	0.23
SAWS-ASPECS flare + CME (CACTUS1) 50%	-1.24	1.24	-0.04
SAWS-ASPECS flare + CME (CACTUS1) 90%	-0.73	1.10	-0.09
SAWS-ASPECS flare + CME (CACTUS2) 50%	-0.92	0.92	-0.10
SAWS-ASPECS flare + CME (CACTUS2) 90%	-0.48	1.22	-0.19
SAWS-ASPECS flare + CME (M2M) 50%	-0.69	0.69	0.22
SAWS-ASPECS flare + CME (M2M) 90%	-0.15	0.48	0.19
SAWS-ASPECS flare + CME (SOHO) 50%	-0.16	0.42	0.47
SAWS-ASPECS flare + CME (SOHO) 90%	0.64	0.64	0.48
SAWS-ASPECS flare + CME (CACTUS1) electrons 50%	-1.21	1.21	0.04
SAWS-ASPECS flare + CME (CACTUS1) electrons 90%	-0.73	1.12	0.04
SAWS-ASPECS flare + CME (CACTUS2) electrons 50%	-0.92	0.92	-0.10
SAWS-ASPECS flare + CME (CACTUS2) electrons 90%	-0.48	1.12	-0.14
SAWS-ASPECS flare + CME (M2M) electrons 50%	-0.69	0.69	0.22
SAWS-ASPECS flare + CME (M2M) electrons 90%	-0.16	0.48	0.19
SAWS-ASPECS flare + CME (SOHO) electrons 50%	-0.15	0.42	0.49
SAWS-ASPECS flare + CME (SOHO) electrons 90%	0.63	0.87	0.49
SAWS-ASPECS flare electrons 50%	-1.01	1.01	0.13
SAWS-ASPECS flare electrons 90%	-0.05	0.65	0.30

Table 6.37: Onset peak metrics for SAWS-ASPECS ≥100 MeV SEPVAL predictions. The metrics include Median Log Error (MedLE), Median Absolute Log Error (MedALE), and Spearman Correlation Coefficient (SCC).

Model	MedLE	MedALE	SCC
SAWS-ASPECS CME (CACTUS1) 50%	-2.13	2.13	0.20
SAWS-ASPECS CME (CACTUS1) 90%	-2.09	2.09	0.21
SAWS-ASPECS CME (CACTUS2) 50%	-2.13	2.13	0.20
SAWS-ASPECS CME (CACTUS2) 90%	-2.09	2.09	0.21
SAWS-ASPECS CME (M2M) 50%	-2.06	2.06	0.17
SAWS-ASPECS CME (M2M) 90%	-2.01	2.01	0.18
SAWS-ASPECS CME (SOHO) 50%	-2.17	2.17	0.21
SAWS-ASPECS CME (SOHO) 90%	-2.12	2.12	0.23
SAWS-ASPECS flare 50%	-0.67	1.21	0.40
SAWS-ASPECS flare 90%	0.96	1.62	0.40
SAWS-ASPECS flare + CME (CACTUS1) 50%	-1.27	1.27	0.51
SAWS-ASPECS flare + CME (CACTUS1) 90%	-0.29	0.95	0.53
SAWS-ASPECS flare + CME (CACTUS2) 50%	-0.25	0.82	0.38
SAWS-ASPECS flare + CME (CACTUS2) 90%	0.88	1.33	0.38
SAWS-ASPECS flare + CME (M2M) 50%	-0.22	0.62	0.61
SAWS-ASPECS flare + CME (M2M) 90%	0.76	1.43	0.61
SAWS-ASPECS flare + CME (SOHO) 50%	0.55	0.88	0.48
SAWS-ASPECS flare + CME (SOHO) 90%	1.90	2.05	0.48
SAWS-ASPECS flare + CME (CACTUS1) electrons 50%	-1.13	1.13	0.49
SAWS-ASPECS flare + CME (CACTUS1) electrons 90%	-0.27	0.95	0.50
SAWS-ASPECS flare + CME (CACTUS2) electrons 50%	-0.26	0.81	0.36
SAWS-ASPECS flare + CME (CACTUS2) electrons 90%	0.88	1.39	0.34
SAWS-ASPECS flare + CME (M2M) electrons 50%	-0.22	0.52	0.68
SAWS-ASPECS flare + CME (M2M) electrons 90%	0.76	1.37	0.65
SAWS-ASPECS flare + CME (SOHO) electrons 50%	0.58	0.84	0.49
SAWS-ASPECS flare + CME (SOHO) electrons 90%	1.93	2.05	0.49
SAWS-ASPECS flare electrons 50%	-0.81	1.21	0.36
SAWS-ASPECS flare electrons 90%	0.66	1.59	0.37

Table 6.38: Max flux metrics for SAWS-ASPECS $\geq$10 MeV SEPVAL predictions. The metrics include Median Log Error (MedLE), Median Absolute Log Error (MedALE), and Spearman Correlation Coefficient (SCC).

Model	MedLE	MedALE	SCC
SAWS-ASPECS CME (CACTUS1) 50%	-1.76	1.76	-0.28
SAWS-ASPECS CME (CACTUS1) 90%	-1.66	1.66	-0.23
SAWS-ASPECS CME (CACTUS2) 50%	-1.76	1.76	-0.28
SAWS-ASPECS CME (CACTUS2) 90%	-1.66	1.66	-0.23
SAWS-ASPECS CME (M2M) 50%	-1.86	1.86	-0.21
SAWS-ASPECS CME (M2M) 90%	-1.77	1.77	-0.31
SAWS-ASPECS CME (SOHO) 50%	-1.86	1.86	-0.09
SAWS-ASPECS CME (SOHO) 90%	-1.77	1.77	-0.28
SAWS-ASPECS flare 50%	-0.79	0.96	0.03
SAWS-ASPECS flare 90%	0.15	0.60	0.26
SAWS-ASPECS flare + CME (CACTUS1) 50%	-1.24	1.24	-0.04
SAWS-ASPECS flare + CME (CACTUS1) 90%	-0.73	1.06	-0.09
SAWS-ASPECS flare + CME (CACTUS2) 50%	-0.92	0.92	-0.10
SAWS-ASPECS flare + CME (CACTUS2) 90%	-0.48	1.22	-0.19
SAWS-ASPECS flare + CME (M2M) 50%	-0.73	0.73	0.22
SAWS-ASPECS flare + CME (M2M) 90%	-0.15	0.48	0.19
SAWS-ASPECS flare + CME (SOHO) 50%	-0.21	0.42	0.45
SAWS-ASPECS flare + CME (SOHO) 90%	0.64	0.64	0.46
SAWS-ASPECS flare + CME (CACTUS1) electrons 50%	-1.21	1.21	0.04
SAWS-ASPECS flare + CME (CACTUS1) electrons 90%	-0.73	1.12	0.04
SAWS-ASPECS flare + CME (CACTUS2) electrons 50%	-0.92	0.92	-0.10
SAWS-ASPECS flare + CME (CACTUS2) electrons 90%	-0.48	1.12	-0.14
SAWS-ASPECS flare + CME (M2M) electrons 50%	-0.73	0.73	0.22
SAWS-ASPECS flare + CME (M2M) electrons 90%	-0.16	0.48	0.19
SAWS-ASPECS flare + CME (SOHO) electrons 50%	-0.20	0.42	0.49
SAWS-ASPECS flare + CME (SOHO) electrons 90%	0.63	0.87	0.49
SAWS-ASPECS flare electrons 50%	-1.06	1.06	0.17
SAWS-ASPECS flare electrons 90%	-0.10	0.68	0.31

Table 6.39: Max flux metrics for SAWS-ASPECS $\geq$100 MeV SEPVAL predictions. The metrics include Median Log Error (MedLE), Median Absolute Log Error (MedALE), and Spearman Correlation Coefficient (SCC).

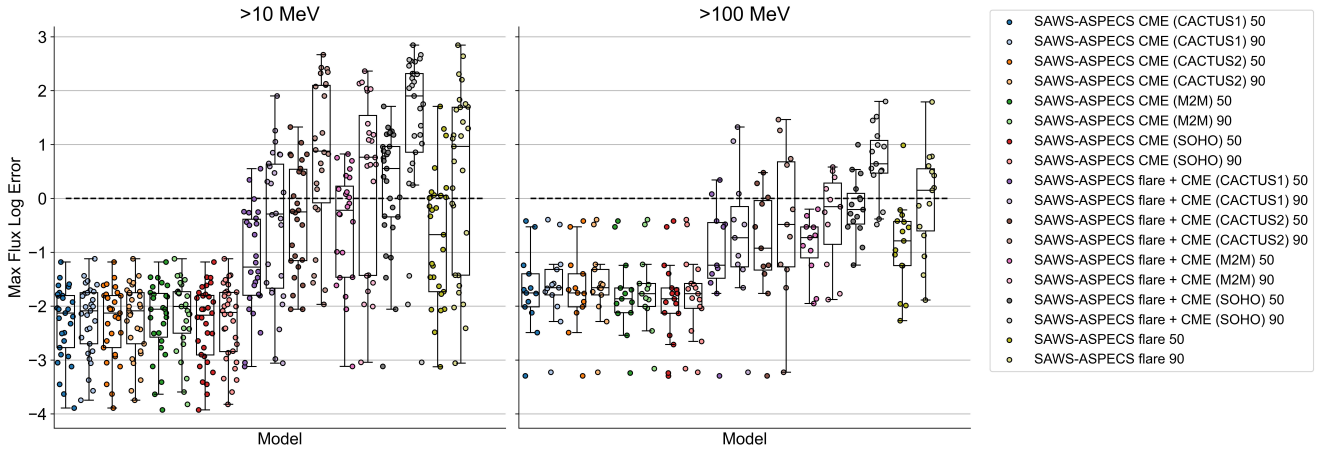


Figure 6.22: Box plots for SAWS-ASPECS ≥10 MeV and ≥100 MeV SEPVAL max flux predictions.

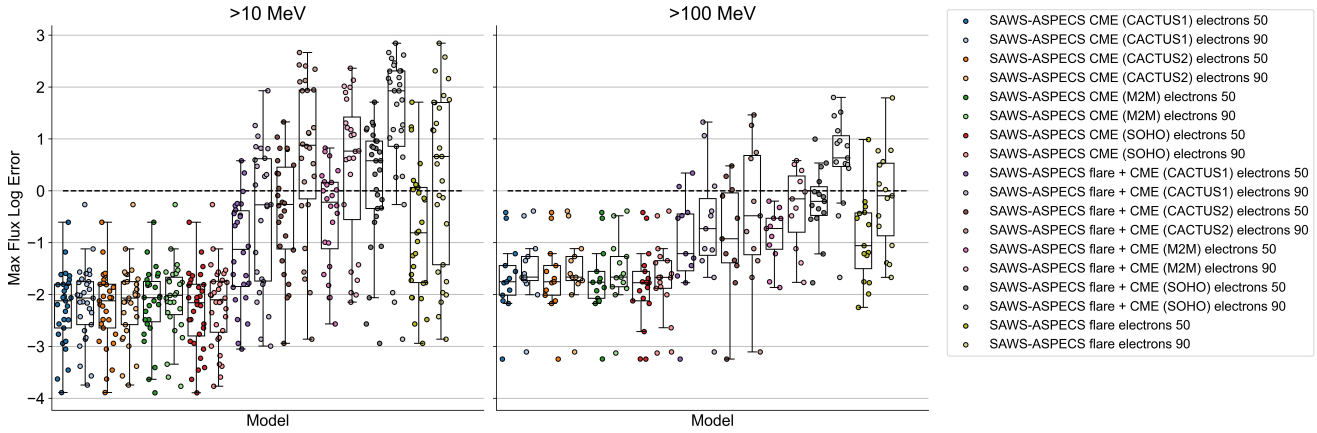


Figure 6.23: Box plots for SAWS-ASPECS ≥10 MeV and ≥100 MeV SEPVAL max flux predictions for variants that include electron flux.

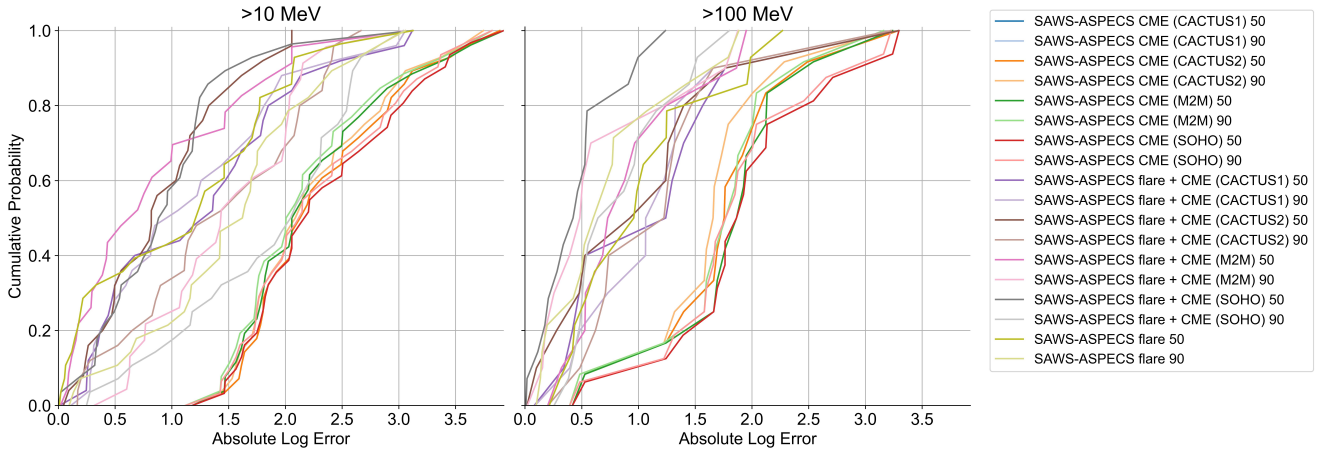


Figure 6.24: Cumulative Distribution Function (CDF) plots for SAWS-ASPECS ≥10 MeV and ≥100 MeV SEPVAL onset peak predictions.

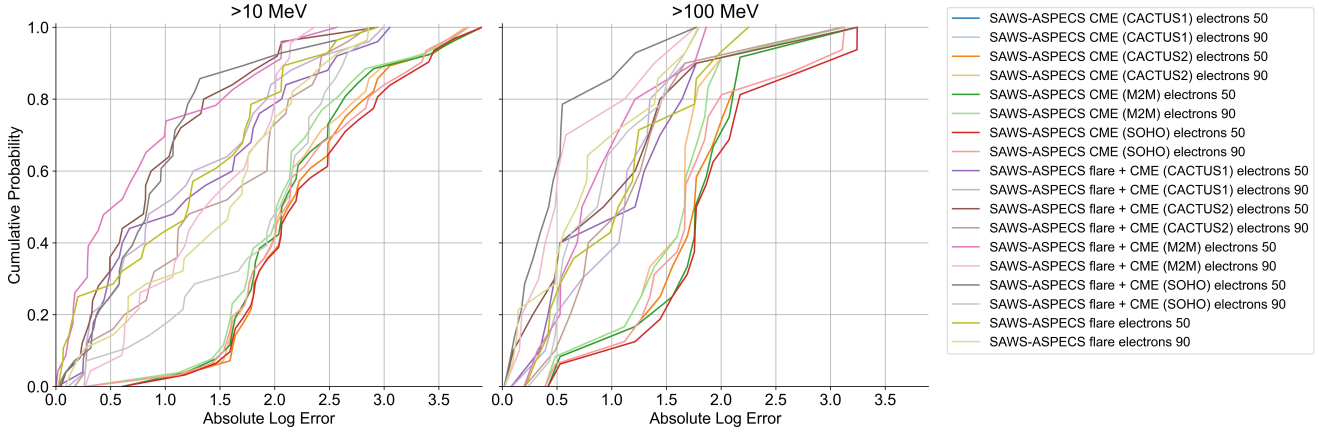


Figure 6.25: Cumulative Distribution Function (CDF) plots for SAWS-ASPECS ≥10 MeV and ≥100 MeV SEPVAL onset peak predictions for variants including electron flux.

6.7 SEPSTER

Solar Energetic Particle STEReo (SEPSTER) is an empirical model, developed by Ian Richardson, that is based on a relationship between CME speed, magnetic connection between the observer and the Sun, and the peak SEP intensity of protons at that observer (Richardson et al., 2014, 2018). This empirical relationship was derived from observations for 25 SEP events observed at multiple spacecraft and fit to a Gaussian function of the source connectivity angle. The relationship is:

$$I_p \sim 0.013 \exp\left(0.0036V - \frac{\phi_{\mathrm{CA}}^2}{2\sigma^2}\right). \tag{6.1}$$

Here, I_p is the peak intensity of 14–24 MeV protons in (Mev s cm^2 sr)$^{-1}$, V is the CME speed in km/s, ϕ_{CA} is the connection angle in degrees, and σ is the average value of the width of the Gaussian functions fitted to the SEP events (43°). The connection angle is calculated as

$$\phi_{\mathrm{CA}} = \frac{1.5 \times 10^8}{u} \frac{360}{86400 \times 27.25} - \phi_{\mathrm{CME}} \tag{6.2}$$

where ϕ_{CA} is the connection angle in degrees, 1 AU is $\sim 1.5 \times 10^8$ km, u is the solar wind speed in km/s, 27.25 days is the Carrington rotation period, and ϕ_{CME} is the longitude of the CME direction with respect to the Sun-Earth line. An alternate version of the connection angle is used in the WSA-ENLIL version of SEPSTER, where the magnetohydrodynamic (MHD) solar wind model provides an angle determined by its coronal mapping. Lastly, to extrapolate the peak intensity I_p to the standard GOES energy ranges, the following factors are used for the >10 MeV, > 30 MeV, > 50 MeV, and >100 MeV, respectively: $\sim$ 20 I_p, $\sim 2I_p$, $\sim$ I_p, and $\sim$ 0.2 I_p. From these, an All Clear binary and peak flux prediction are provided.

SEPSTER (Parker Spiral)

Characteristic	> 10 MeV	> 100 MeV
First Forecast	2020/05/13	2020/05/13
Last Forecast	2024/11/03	2024/11/03
N Days	1665	1665
N Forecast Days	1250	1250
N SEP Days	31	7
Forecast Cadence	Triggered	Triggered
Prediction Window	Varies (hours)	Varies (hours)
N forecasts	3899	3899
N matched w/events	61	15
Imbalance (raw)	62.92	258.93
Imbalance (days)	39.32	177.57

Table 6.40: SEPSTER (Parker Spiral) validation characteristics table.

As a caveat to interpreting this model, SEPSTER was originally tuned for SOHO and STEREO predictions, where a scaling factor is applied to match to the GOES

SEPVAL

> 10 MeV

	Observed Yes	No	Sum
Pred. Yes	20	5	25
Pred. No	11	25	36
Sum	31	30	61

> 100 MeV

	Observed Yes	No	Sum
Pred. Yes	7	2	9
Pred. No	9	44	53
Sum	16	46	62

Scoreboard

> 10 MeV

	Observed Yes	No	Sum
Pred. Yes	38	68	106
Pred. No	23	3634	3657
Sum	61	3702	3763

> 100 MeV

	Observed Yes	No	Sum
Pred. Yes	1	13	14
Pred. No	11	3042	3053
Sum	12	3055	3067

Table 6.41: SEPSTER (Parker Spiral) contingency tables.

integral channels. Due to a singular scaling factor for each integral channel, there are some expected discrepancies between what is predicted and observed. Additionally, equation 6.1 was based on 2D plane-of-sky CME speeds from the Coordinated Data Analysis Workshop (CDAW) CME catalog, whereas SEPSTER uses 3D CME parameters from the DONKI catalog in real time and for the SEPVAL challenge. It is also worth noting that SEPSTER predicts the peak of a potential SEP event for all input CMEs, while climatologically a much smaller subset of CMEs actually produce an SEP event.

SEPSTER provided forecasts for 62 of the SEPVAL challenge periods, running in historical mode and using the refit CME parameters provided by the M2M office. Here we focuse on in its Parker Spiral 'mode', since the WSA-ENLIL version produced very similar results. To determine its binary All Clear forecasts, SEPSTER compares the peak flux prediction to the operational thresholds of 10 pfu (for >10 MeV, SPE) and 1 pfu (for >100 MeV, ESPE). Table 6.41 reports the contingency tables for SEPVAL and the SEP Scoreboards with the resulting metrics in Table 6.42.

For the balanced SEPVAL dataset, SEPSTER shows some skill in differentiating the conditions that cause SEP enhancements above threshold in the >10 MeV channel versus those that do not. This is demonstrated by a Hit Rate of 64.5% paired with a fairly low False Alarm Rate of 16.7% and False Alarm Ratio of 20.0%. When comparing SEPSTER outputs to random chance, the HSS of 0.48 shows medium skill compared to other SEPVAL participants (see Figure 5.5). For ESPE enhancements, SEPSTER has a much lower Hit Rate of 43.8% but also a low False Alarm Rate of 4.3% and False Alarm Ratio of 22%, indicating that SEPSTER underpredicts the intensity of the >100 MeV channel. Despite the decrease in Hit Rate, the correspondingly low False Alarm Rate and Ratio results in a relatively high HSS score of 0.46 compared to other models (see Figure 5.7). When looking at the unbalanced, climatological dataset as the model is running on the SEP Scoreboards, SEPSTER's performance shows a decrease in skill. For the >10 MeV channel the hit rate is consistent at 62.3%, while there is a large decrease for the >100 MeV

SEPSTER All Clear	SEPVAL		Scoreboard	
	>10 MeV ($N=61$)	>100 MeV ($N=62$)	>10 MeV ($N=3763$)	>100 MeV ($N=3067$)
Percent Correct	0.74	0.82	0.98	0.992
Hit Rate	0.65	0.44	0.62	0.08
False Alarm Rate	0.17	0.043	0.018	0.0043
False Alarm Ratio	0.20	0.22	0.64	0.93
Bias	0.81	0.44	1.74	1.17
Threat Score	0.56	0.39	0.29	0.04
HSS	0.48	0.46	0.44	0.07
TSS	0.48	0.39	0.60	0.08

Table 6.42: SEPSTER (Parker Spiral) All Clear metrics.

channel to 8.3%. False alarms, represented by the False Alarm Ratio, increase to 65% and 93% for >10 MeV and >100 MeV, respectively, which are both much larger than the SEPVAL values, meaning that when the model gives a "yes" forecast (All Clear False) more than half of the time that is incorrect. Despite this, the FAR values are still low, meaning that the model captures most SEP quiet times on the Scoreboards. The HSS values show the model's skill compared to random chance, and for the >10 MeV channel SEPSTER shows similar skill to SEPVAL with a value of 0.44. In fact, an HSS of 0.44 is the second highest score for all models on the SEP Scoreboards, indicating excellent relative performance for >10 MeV in real time. However, for the >100 MeV channel, the value of 0.07 shows no skill compared to random chance, this is mainly due to the model correctly predicting only 1 ESPE, resulting in a Hit Rate in the single digit percents.

As its main output, SEPSTER gives a peak intensity prediction, which is most comparable to the onset peak intensity of an SEP event, rather than the maximum peak flux which could be related to the arrival of a CME shock at an observer resulting in an ESP enhancement. ESPs are typically associated with particle energies below 100 MeV, so for the >100 MeV energy channel, the onset peak and maximum flux are typically the same or closer in value. SEPSTER's peak flux predictions were compared to both the observed onset peak and the maximum flux despite this caveat. The onset peak metrics are shown in Table 6.43 and the max flux metrics are shown in Table 6.44. For the >10 MeV channel in the SEPVAL challenge and on the Scoreboards, SEPSTER shows a bias to underpredict onset peak by nearly a factor of 3 shown in the median log error ($log_{10}(3) = 0.47$) and with an accuracy within a factor of 5 ($log_{10}(5) \sim 0.69$). As expected from the caveat, the performance for max flux is slightly worse across the metrics, but the predictions are still typically within an order of magnitude. This shows that SEPSTER has some skill in predicting the magnitude of SEP events.

For the >100 MeV energy channel, SEPSTER has a decrease in performance across all metrics as seen in Tables 6.43 and 6.44. For the SEPVAL set, we now see a bias in the onset peak predictions of an order of magnitude (MedLE $= -0.961$) and accuracy decrease by an order of magnitude (MedALE $= 0.961$). On the Score-

SEPSTER Onset Peak Flux	SEPVAL > 10 MeV ($N = 31$)	SEPVAL > 100 MeV ($N = 16$)	Scoreboard > 10 MeV ($N = 57$)	Scoreboard > 100 MeV ($N = 15$)
Percent within a factor of 10	74	50	64	20
Percent within a factor of 2	36	13	50	0
Median Log Error	−0.43	−0.92	−0.36	−1.5
Median Absolute Log Error	0.65	0.92	0.59	1.5
Spearman Correlation Coefficient	0.59	0.28	0.48	0.26

Table 6.43: SEPSTER (Parker Spiral) onset peak metrics.

SEPSTER Max Peak Flux	SEPVAL > 10 MeV ($N = 31$)	SEPVAL > 100 MeV ($N = 16$)	Scoreboard > 10 MeV ($N = 57$)	Scoreboard > 100 MeV ($N = 15$)
Percent within a factor of 10	71	50	60	20
Percent within a factor of 2	32	13	24	0
Median Log Error	−0.71	−0.96	−0.50	−1.6
Median Absolute Log Error	0.81	0.96	0.63	1.6
Spearman Correlation Coefficient	0.51	0.28	0.40	0.34

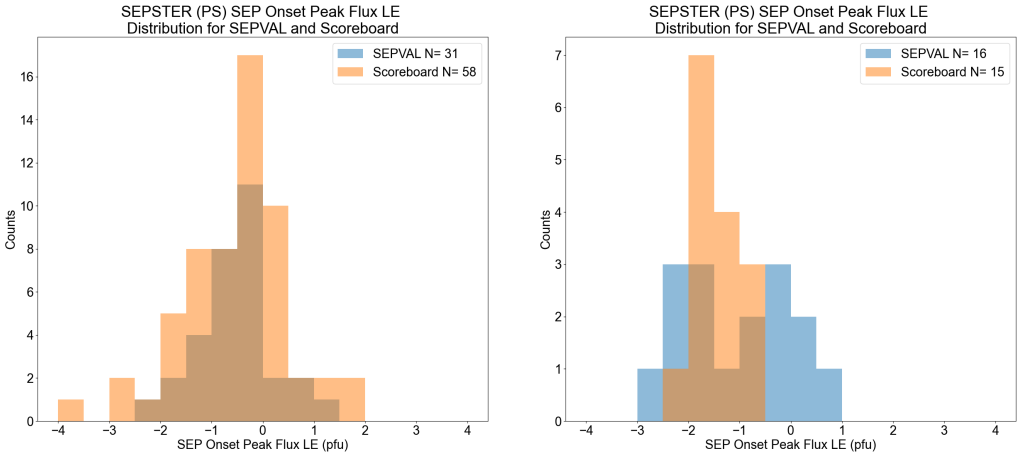
Table 6.44: SEPSTER (Parker Spiral) max flux metrics.

boards, these values are over an order of magnitude at 1.5, which explains the previously noted tendency for SEPSTER to miss ESPE events in real time. The metrics are nearly identical for the max flux comparison for this energy channel.

From the correlation coefficients in Table 6.43 and 6.44, we can see that SEPSTER has moderate skill in differentiating between large and small SEP events (predicting a larger flux for larger events and a smaller flux for smaller events), with a Spearman coefficient of 0.59 for onset peak and 0.51 for maximum flux for the >10 MeV protons. The correlation is slightly lower for onset peak (0.48) and max flux (0.40) Scoreboard predictions. For >100 MeV, the Spearman correlation coefficients are much lower, indicating that the event-to-event trends are not well captured.

It is useful to quantify how often a model's peak forecast is within an order of magnitude of the observed peak. Figure 6.26 shows the distributions of the log error in the onset peak flux for the two operational energy channels and for each dataset. For the >10 MeV channel, for each dataset (SEPVAL and the SEP Scoreboards) the distribution is centered at the bin located from −0.5 to 0 and a longer tail in the negative direction, showing SEPSTER's bias towards underprediction, but with most predictions within an order of magnitude. For the >100 MeV energy channel, the SEPVAL data shows a bimodal distribution, with one peak at LE = -2 and another at LE = -0.5. Using the VIVID visualization tool, Figure 6.27 shows that for progressively slower CME speeds, SEPSTER gives a worse prediction for the onset peak flux. The outlier point with a -2.09 log error for a CME with a speed just above 2000 km/s is due to an Eastern CME with a longitude of -29°, which as we will see is an expected incorrect prediction due to SEPSTER's empirical relationship.

SEPSTER's empirical relationship invokes a dependency on CME longitude that



(a) >10 MeV (b) >100 MeV

Figure 6.26: Histogram of onset peak log error for SEPSTER (Parker Spiral) for SEPVAL (blue) and the SEP Scoreboards (orange) for >10 MeV (left) and >100 MeV (right).

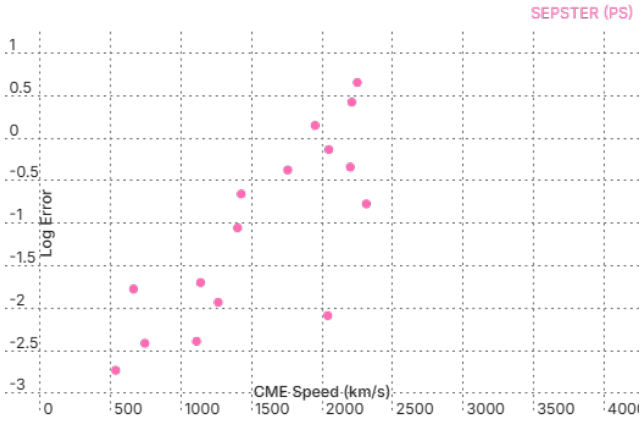


Figure 6.27: Log error of SEPSTER (Parker Spiral) predicted onset peak flux for >100 MeV energy channel dependency on source CME speed.

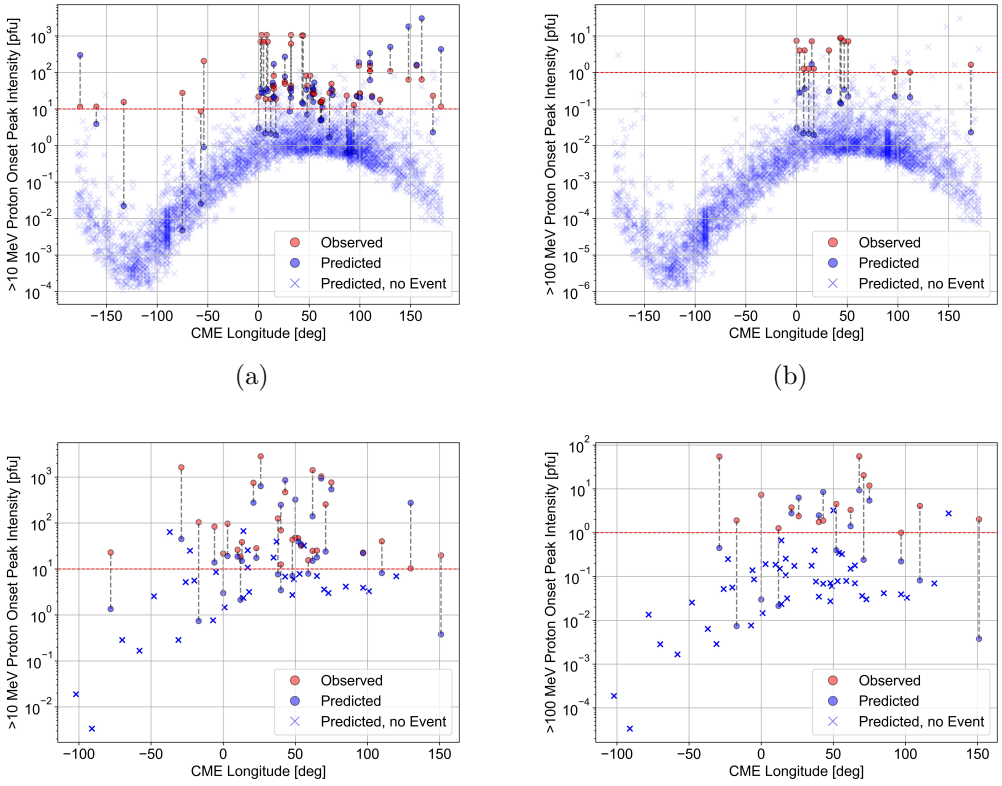


Figure 6.28: SEPSTER (Parker Spiral) onset peak performance versus CME longitude for all forecasts made to the SEP Scoreboards and the SEPVAL dataset. The observed onset peak (red) and predicted onset peak (blue), connected with dotted lines, show the differences as a function of CME longitude. Blue crosses are predicted values with no observed SEP event. These plots are a recreation of a plot from Richardson et al. (2018). a) >10 MeV SEP Scoreboards b) >100 MeV SEP Scoreboards c) >10 MeV SEPVAL d) >100 MeV SEPVAL

is not seen in the observed SEP peak values in the SEPVAL challenge periods and on the SEP Scoreboards, shown in Figure 6.28. Figures 6.28a and 6.28b are a recreation of a similar plot from the original paper, but now with data taken from the SEP Scoreboards. Figure 6.28a shows the observed and predicted onset peak flux for >10 MeV protons, where it is seen that the peak is overpredicted for Western CMEs and underpredicts for Eastern CMEs. This plot is recreated for the SEPVAL data in Figures 6.28c and 6.28d, where the observations do not show as strong of a dependence on source longitude as the SEPSTER predictions do. Extrapolation from this shows that for Eastern CMEs SEPSTER underpredicts these CMEs, since all of its predictions for CMEs further east than $-50°$ are below 10 pfu. For the SEP Scoreboards, this behavior results in misses for 6 SEP events, albeit most of those >10 MeV events were small. For SEPVAL, one >10 MeV event east of $-50°$ was missed.

Interestingly, Figures 6.28a and 6.28b provide insight into SEPSTER's underprediction and poor performance in All Clear for >100 MeV events compared to its >10 MeV performance. In Figure 6.28a, the >10 MeV predictions are sometimes higher and sometimes lower than the observed peak values, as also shown in the histogram in Figure 6.26 (left), with many predictions above the 10 pfu threshold (red dashed line). Since SEPSTER applies a single scaling factor to generate >100 MeV predictions, every prediction is exactly mirrored in Figure 6.28b, however all of the values are underpredictions compared to observed peaks and the entire curve with CME longitude is lower with respect to the 1 pfu threshold (red dashed line) compared to the >10 MeV curve. Optimizing the >100 MeV scaling factor using both the error in the flux and the All Clear metrics may improve SEPSTER's >100 MeV peak flux skill.

AWT for SEPSTER is primarily driven by the time it takes for SOHO-Large Angle and Spectrometric Coronagraph Experiment (LASCO) imagery to be downlinked and any observed CME to be measured by the M2M office and uploaded to DONKI. For the most prompt SEP events, typically driven by Western CMEs where there is favorable magnetic connectivity, these CME measurements are often available only after the SEP event has started. This is seen in the median AWT to SEP start time of -37 minutes, shown in Table 6.45. The negative warning time means forecasts are issued after the observed timing (in this case SEP start time) and half of SEPSTER forecasts are typically available to the end user 36 minutes or more *after* an observed >10 MeV event has started. However, since SEPSTER also provides a prediction for SEP onset peak, we also can measure the advance warning for the peak, where SEPSTER provides 4 hours of warning.

A summary of SEPSTER's validation performance is below:

- All Clear >10 MeV: SEPSTER has a strong resilience to the balance of the dataset, where its all clear metrics don't change much from the SEPVAL to the SEP Scoreboard. Consistent Hit Rate within $\sim 60\%$ and fairly low False Alarm Rate and good HSS scores.

- All Clear >100 MeV: SEPSTER shows good skill for the SEPVAL dataset, with a bias towards underpredicting events. For the Scoreboard, a Hit Rate

SEPSTER (Parker Spiral) Advance Warning Time (AWT)	Scoreboard > 10 MeV ($N = 15$)	> 100 MeV ($N = 0$)
AWT to Observed SEP Start Time		
Median	−37 min	-
Worst	−10.9 hr	-
Best	6.8 hr	-
AWT to Observed SEP Onset Peak Time		
Median	4.1 hr	-
Worst	−4.4 hr	-
Best	15.8 hr	-

Table 6.45: Advance Warning Time for SEPSTER (Parker Spiral) on the SEP Scoreboard for >10 and >100 MeV.

less than 10%, balanced with False Alarm Rate less than 1% shows no skill in predicting ESPE events in real time.

- Onset Peak >10 MeV: Most SEPSTER predictions are within an order of magnitude of the observed onset peak ($\sim$ 2/3 of events), with a slight bias towards underpredicting by a factor of 3.
- Onset Peak >100 MeV: SEPSTER underpredicts all Scoreboard ESPE events and shows little skill for the SEPVAL dataset (only 50% of predictions are within an order of magnitude).
- Skill in >100 MeV SEPSTER predictions might be improved by adjusting the scaling factor to optimize across All Clear and onset peak flux metrics.
- Advance Warning Time: SEPSTER provides no advance warning for the start of an SEP event, with a forecast to be expected to be issued a median of 37 minutes after event start, however a median of $\sim$ 4 hours of advance warning are provided for the onset peak. Not enough data for ESPE events.

6.8 SEPSTER2D

Developed from the SEPSTER framework, SEPSTER2D has similar assumptions that SEP intensity can be predicted from CME speed and angle between the CME and observing spacecraft. Where this model differs is that its empirical relationship was derived using multi-point spacecraft measurements of GOES-13/15, STEREO A and B, and PAMELA. Some of the data quality issues in GOES data (see Section 3.6) were addressed by performing a background subtraction and applying the effective energy bins derived by Sandberg et al. (2014) to GOES P2-P5 proton energy channels and effective energy bins derived by Bruno (2017) the GOES P6-P11 channels. The Sandberg et al. (2014) effective energies are based on a cross-calibration of GOES with Interplanetary Monitoring Platform (IMP)-8/Goddard Medium Energy (GME) while Bruno (2017) performed a cross-calibration with Payload for Antimatter Matter Exploration and Light-nuclei Astrophysics (PAMELA). It should be noted that the model was developed using these calibrated data while the validation performed here is with respect to the GOES integral fluxes provided by NOAA. The background-subtraction and application of calibrated effective energies has the largest effect on the high-energy part of the SEP spectrum, causing the >100 MeV SEPSTER2D predictions to differ significantly from the NOAA-provided GOES integral fluxes, which tend to have high backgrounds in the higher energy bins.

SEPSTER2D's SEP intensity distribution is given as a function of energy E and spherical distance from the location of peak SEP intensity δ:

$$\Phi(E,\delta) = \Psi_{CME}(E)\exp\left(\Lambda_{CME}V_{CME}\right)G(E,\delta) \tag{6.3}$$

where Ψ_{CME} and Λ_{CME} account for the energy-dependence of the intensity, V_{CME} is the CME speed, and $G(E,\delta)$ is a Gaussian distribution accounting for both latitudinal and longitudinal magnetic connectivity. The location of the SEP peak is assumed to coincide with the estimated CME latitudinal angle based on coronagraph imagery. Similar to SEPSTER, the footpoint of the Interplanetary Magnetic Field (IMF) field line between the CME and the observer is based on a Parker spiral, using the solar wind speed.

A limitation of SEPSTER2D is that it is known to overpredict for slow CMEs and narrow SEP events (Whitman et al., 2023), and that the intensity prediction for energies above 130 MeV are based on spectral extrapolations with large uncertainties. SEPSTER2D developed its empirical relationships using a "peak" spectrum in which the maximum flux value for each differential energy channel was combined into a single spectrum, fit with a Band function, and then integrated to create an integral flux value representing the SEP maximum (private communication with the developer). The construction of this asynchronous peak spectrum may explain the tendency to overpredict for certain cases.

SEPSTER2D has been running in real time on the SEP Scoreboard since December 2020, where it has produced 1103 forecasts and for the SEPVAL challenge it provided forecasts for 60 of the challenge periods. The contingency tables for these two datasets are in Table 6.47 and the respective metrics are in Table 6.48.

	SEPSTER2D	
Characteristic	> 10 MeV	> 100 MeV
First Forecast	2021/06	2021/06
Last Forecast	2024/12/30	2024/12/30
N Days	2674	2674
N Forecast Days	562	562
N SEP Days	31	9
Forecast Cadence	Triggered	Triggered
Prediction Window	Varies (hours)	Varies (hours)
N forecasts	1103	1103
N matched w/events	63	17
Imbalance (raw)	16.51	63.88
Imbalance (days)	17.13	61.44

Table 6.46: SEPSTER2D validation characteristics table.

SEPSTER2D shows resilience to the balance of the dataset, with many of its All Clear metrics having similar values across the two datasets. For SEPVAL and the SEP Scoreboards, respectively, the >10 MeV channel predictions result in a percent correct of 71.2% (69.1%), Hit Rate of 90% (94%), and False Alarm Rate of 48% (33%). The two datasets begin to diverge with False Alarm Ratios of 34% and 84%, Threat Score of 0.61 and 0.16, HSS of 0.42 and 0.19, and TSS of 0.42 and 0.61. The False Alarm Rate reports the number of false alarms with respect to all negative periods, meaning that the metric is significantly impacted by the large number of correct negatives. The False Alarm Ratio compares the number of false alarms to all "yes" forecasts, providing a ratio that takes into account the (much smaller) number of hits. In the Scoreboard dataset, SEPSTER2D tends to overpredict the occurrence of SEP events (over 300 false alarms) which causes the False Alarm Ratio to increase. The TSS, however, is higher for the Scoreboard dataset despite these false alarms due to the large number of correct negatives in the denominator which suppresses the impact of these false alarms. For the >100 MeV channel, SEPSTER2D tends to underpredict the occurrence of ESPE events, seen in the Hit Rate of 50% for SEPVAL (18% for Scoreboards), with low False Alarm Rate 11% (3.2% for the Scoreboards). The large False Alarm Ratio for the Scoreboards (92%) is due to there being more than 10x times false alarms compared to hits, which also affects the HSS and TSS negatively (as do the number of misses for this dataset).

SEPSTER2D was trained to predict the maximum flux of SEP events, so we focus the discussion on those metrics in Table 6.49 (but the onset peak flux are still presented in Table 6.50 for comparison to SEPSTER). Continuing the trend from the All Clear metrics, SEPSTER2D shows some resilience to the balance of the dataset. The log error distributions for both SEPVAL and the Scoreboards for the >10 MeV are centered around 0 with median values of -0.04 and 0.07, showing no bias for under- or overpredicting. As for accuracy, median absolute log error shows that, for both datasets, SEPSTER2D is inaccurate by about a factor of 3 (MALE =

SEPVAL

> 10 MeV	Observed Yes	Observed No	Sum
Pred. Yes	27	14	41
Pred. No	3	15	18
Sum	30	29	59

> 100 MeV	Observed Yes	Observed No	Sum
Pred. Yes	8	5	13
Pred. No	8	39	47
Sum	16	44	60

Scoreboard

> 10 MeV	Observed Yes	Observed No	Sum
Pred. Yes	59	313	372
Pred. No	4	649	653
Sum	63	962	1025

> 100 MeV	Observed Yes	Observed No	Sum
Pred. Yes	3	35	38
Pred. No	14	1043	1057
Sum	17	1078	1095

Table 6.47: SEPSTER2D contingency tables.

SEPSTER2D All Clear	SEPVAL > 10 MeV ($N = 59$)	SEPVAL > 100 MeV ($N = 60$)	Scoreboard > 10 MeV ($N = 1025$)	Scoreboard > 100 MeV ($N = 1095$)
Percent Correct	0.71	0.78	0.69	0.96
Hit Rate	0.90	0.50	0.94	0.18
False Alarm Rate	0.48	0.12	0.33	0.032
False Alarm Ratio	0.34	0.39	0.84	0.92
Bias	1.37	0.56	5.90	2.24
Threat Score	0.61	0.38	0.14	0.06
HSS	0.42	0.41	0.19	0.09
TSS	0.42	0.39	0.61	0.14

Table 6.48: SEPSTER2D All Clear Metrics

0.50 and 0.57). Accuracy is also represented in the log error histogram (Figure 6.29a, where over 80% of forecasts are within a factor of 10 for both datasets, and one-third of the forecasts are within a factor of 2. The plot 6.29b shows that, for the >100 MeV, there is a tendency to underpredict the max peak, with none of the ESPE events having a prediction over the observed value, and both distributions have tails extending in the negative direction. This is also seen in the table where the median log error is -0.91 and -1.12 for SEPVAL and the Scoreboards, respectively.

Similar to SEPSTER, SEPSTER2D's advance warning is limited by the acquisition of SOHO-LASCO coronagraph imagery for CME parameters to be measured. In the case of SEP start times, SEPSTER2D does not provide any advance warning but instead issues forecasts $\sim 0.5 - 1$ hour after the event starts, as shown in Table 6.51. However, it does provide a median AWT of 3.5 hours ahead of the onset peak. The very large "worst AWT" values listed in the table are likely due to extended time periods when a human analyst was not available to measure CME parameters or the model was rerun at a later date after CME parameters were revised.

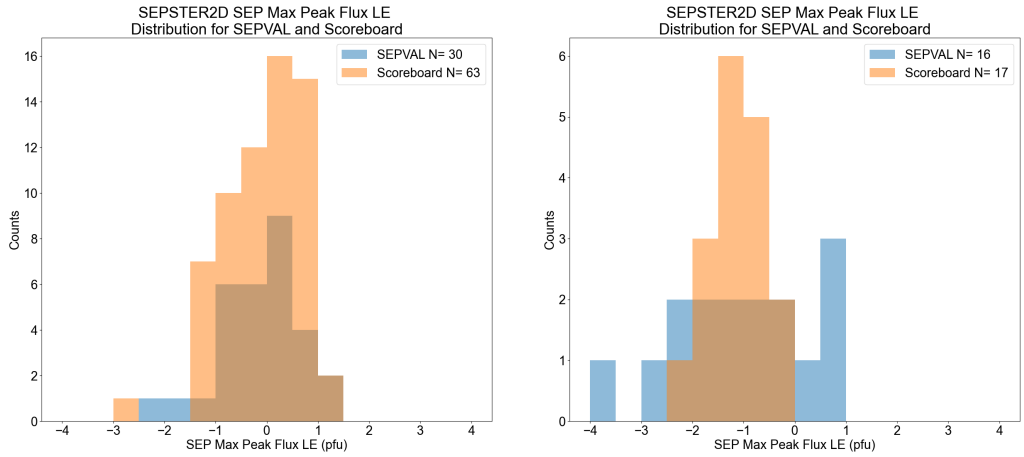
A summary of the validation results for this model is below:

SEPSTER2D Max Peak Flux	SEPVAL		Scoreboard	
	> 10 MeV ($N = 30$)	> 100 MeV ($N = 16$)	> 10 MeV ($N = 63$)	> 100 MeV ($N = 17$)
Percent within a factor of 10	83	50	84	41
Percent within a factor of 2	33	6.3	32	0
Median Log Error	-0.04	-0.91	0.07	-1.12
Median Absolute Log Error	0.50	0.97	0.57	1.12
Spearman Correlation Coefficient	0.59	0.23	0.40	0.46

Table 6.49: SEPSTER2D Max Peak Flux Metrics.

SEPSTER2D Onset Peak Flux	SEPVAL		Scoreboard	
	> 10 MeV ($N = 30$)	> 100 MeV ($N = 16$)	> 10 MeV ($N = 62$)	> 100 MeV ($N = 17$)
Percent within a factor of 10	87	50	78	41
Percent within a factor of 2	33	6.3	29	0.0
Median Log Error	0.10	-0.86	0.26	-1.21
Median Absolute Log Error	0.42	0.97	0.53	1.12
Spearman Correlation Coefficient	0.64	0.21	0.52	0.39

Table 6.50: SEPSTER2D Onset Peak Flux Metrics.



(a) >10 MeV

(b) >100 MeV

Figure 6.29: Histogram of SEPSTER2D max peak log errors for SEPVAL (blue) the SEP Scoreboards (orange) in the >10 MeV (left) and >100 MeV (right) energy channels.

SEPSTER2D Advance Warning Time (AWT)	Scoreboard > 10 MeV ($N = 27$)	> 100 MeV ($N = 1$)
AWT to Observed SEP Start Time		
Median	−33 min	−53 min
Worst	−53.9 hr	-
Best	6.8 hr	-
AWT to Observed SEP Onset Peak Time		
Median	3.5 hr	3.8 hr
Worst	−53.5 hr	-
Best	27.1 hr	-

Table 6.51: Advance Warning Time for SEPSTER2D on the SEP Scoreboard for >10 and >100 MeV.

- All Clear >10 MeV: SEPSTER2D shows skill in correctly predicting events in the >10 MeV energy channel, but the tendency to overpredict results in a corresponding increase in false alarms, causing a significant decrease in skill in real time compared to the historical SEPVAL dataset.
- All Clear >100 MeV: With a low Hit Rate and high False Alarm Ratio, SEPSTER2D shows no skill in predicting the occurrence of ESPE in real time and, with moderate Hit Rate and low False Alarm Rate, shows some skill for the historical periods of the SEPVAL challenge.
- Max Peak Flux >10 MeV: SEPSTER2D shows no bias for under- or over-predicting the maximum peak of SPE events, and most predictions (> 75% for both datasets) are within a factor of 10 and ∼1/3 are within a factor of 2. Correlation coefficients show moderate skill in capturing event-to-event variability.
- Max Peak Flux >100 MeV: SEPSTER2D has a bias to underpredict ESPE typically by nearly an order of magnitude or more, with moderate skill in capturing the max flux within a factor of 10, and no skill within a factor of 2. Only slight skill in capturing event-to-event variability.
- Advance Warning Time: SEPSTER2D's dependence on the acquisition of CME parameters results in typically no advance warning prior to observed threshold crossings, however the model can provide advance warning ahead of the onset and maximum peaks.

6.9 ZEUS+iPATH

The iPATH model is a physics-based SEP model that simulates particle acceleration at CME shocks and the transport of the energetic particles through the inner heliosphere. iPATH is coupled to the MHD model ZEUS (Clarke (1996)) which presents a simplistic solar wind background where a CME can be inserted as a time-dependent perturbation on the inner boundary of the model (0.05 Astronomical Unit (AU)). The CME shock system is modeled as a 2-D onion shell model, where as the CME propagates outward radially new shells are created and old shells are convected and adiabatically expand with the solar wind. Particle distributions are tracked in cells along the shock front and are allowed to diffuse backwards and between cells via parallel and perpendicular diffusion. Particle distributions that travel a certain distance in a single time step escape the shock and their transport is modeled with a focused transport scheme. The focused transport equation is solved with a time-backwards stochastic differential equation where ensemble averages of the many test particle paths give the full particle distribution function at the chosen observer locations. From this distribution, the model then provides time profile outputs for many differential proton channels, which are turned into integral flux time profiles. A full description of this model is available in Hu et al. (2017)

ZEUS+iPATH		
Characteristic	> 10 MeV	> 100 MeV
First Forecast	2023/06/24	2023/06/24
Last Forecast	2024/12/29	2024/12/29
N Days	568	568
N Forecast Days	362	362
N SEP Days	18	5
Forecast Cadence	Triggered	Triggered
Prediction Window	72 hours	72 hours
N forecasts	869	869
N matched w/events	37	12
Imbalance (raw)	22.49	71.4
Imbalance (days)	19.11	71.42

Table 6.52: ZEUS+iPATH validation characteristics table.

To run the model in real time on the SEP Scoreboards, some modifications are made to shorten its computation time. Instead of creating a new solar wind background for each CME input at the time when the CME is injected into the simulation (like other physics-based models may do), in real time, iPATH creates a background solar wind simulation with ZEUS every 8 hours which is then used for all CMEs that occur before the next solar wind background simulation is created. Additionally, a filter is applied that excludes CMEs with speeds less than 450 km/s, which decreases the number of runs in the pipeline that could cause computational pile-ups, and decreases the clutter on the Scoreboard displays. Another real-time addition is the calculation of the suprathermal seed particle spectrum from Deep

SEPVAL

> 10 MeV	Observed Yes	Observed No	Sum
Pred. Yes	20	4	24
Pred. No	11	25	36
Sum	31	29	60

> 100 MeV	Observed Yes	Observed No	Sum
Pred. Yes	8	7	15
Pred. No	7	37	44
Sum	15	44	59

Scoreboard

> 10 MeV	Observed Yes	Observed No	Sum
Pred. Yes	22	49	71
Pred. No	15	729	744
Sum	37	778	815

> 100 MeV	Observed Yes	Observed No	Sum
Pred. Yes	8	53	61
Pred. No	4	802	806
Sum	12	855	867

Table 6.53: ZEUS+iPATH contingency tables.

Space Climate Observatory (DSCOVR) or ACE/EPAM (whichever is available) proton measurements, which is then injected as an initial distribution into each cell along the shock front and then accelerated. This allows the model to try to capture some of the event-to-event variability in the seed particles which has an influence on the predicted magnitude of an event.

ZEUS's simplified approach should be noted as a caveat in how well the model can reproduce the time profile of an SEP event. In real time, iPATH takes the solar wind parameters at 1 AU as input, scales those parameters to the radius of the ZEUS inner boundary, then allows the simulation to reach a steady-state solution using a Parker Spiral. This is an oversimplification of the solar wind, as there are no longitudinal dependencies in the solar wind parameters, which one would get with a more sophisticated solar wind model. Another caveat is the inner boundary is at 0.05 AU, which is above the theorized height where CME shock formation and particle acceleration occurs in the lower corona, meaning the model does not accurately capture the physics occurring during the onset of an SEP event. Lastly, iPATH is only capable of modeling a single CME at a time. During more complex periods where there are multiple CMEs erupting in succession leading to higher chances of particle acceleration leading to SEP events, the model is not able to capture the additional complexity of these interacting CMEs.

As a preface to the validation results in this publication, after receiving the first look community challenge scores during the SEPVAL workshop in 2023, iPATH identified weaknesses in performance and has since been updated with new physics improvements meant to correct some of the problems seen in these metrics. The SEPVAL results reported here reflect the original set of forecasts submitted to the challenge, not including the R2O2R improvements. The model changes were eventually implemented on the SEP Scoreboards, so the Scoreboard contains mixed versions of iPATH, which is true for many of the Scoreboard models.

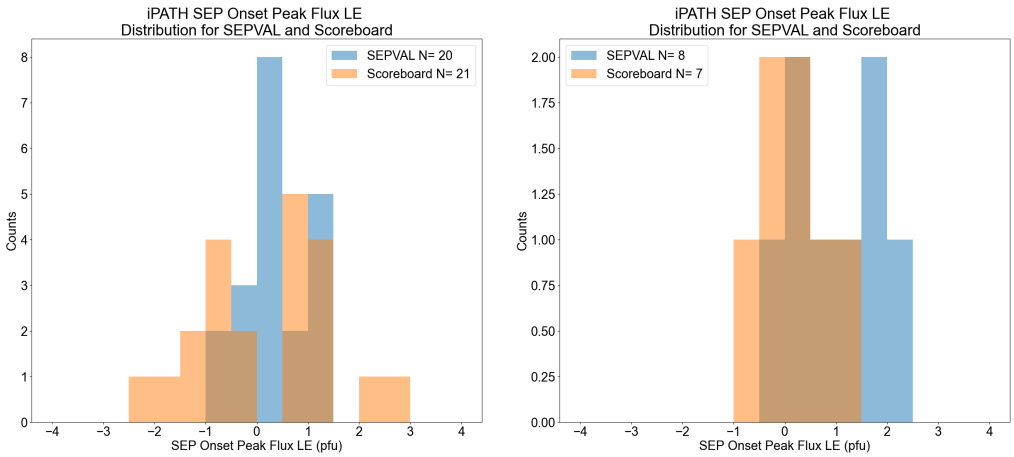
Since iPATH's main output is a time profile, the SEP characteristics are extracted using FetchSEP. All Clear is determined by whether the predicted flux time series rise above SRAG's operational thresholds. Table 6.53 contains the contingency

tables for both datasets and energies, and Table 6.54 contains the corresponding metrics. For the balanced dataset of the SEPVAL challenge, iPATH shows similar performance to the median of the SEPVAL models for both energies. For the >10 MeV (>100 MeV) channel, the percent correct of 75% (77%) and Hit Rate of 65% (53%) show that the model generally correctly predicts the occurrence of SEP events while keeping the False Alarm Rate low at 14% (16%). The >10 MeV predictions have fairly high HSS and TSS of 0.50 and 0.51. The >100 MeV channel, however, has a higher False Alarm Ratio of 0.47 due to the nearly equal number of false alarms to hits. The skill scores drop accordingly with HSS and TSS of 0.37. The real-time Scoreboard dataset also shows large False Alarm Ratios of 69% and 87% for both energies. Combining the False Alarm Ratio with the Hit Rates (59% and 67% respectively) for the Scoreboard shows that, in real time, iPATH has lower skill in detecting the source/cause of SEP events. This is also seen in the TSS, which indicates how well a model differentiates between 'no' and 'yes' events, with values of 0.37 and 0.20. When compared to random chance, however, scores of 0.53 and 0.61 for >10 and >100 MeV, respectively, shows the model has some skill to discriminate between event and non-event periods in real time. Summing up the real-time performance of iPATH, we see that the model tends to forecast many false alarms compared to the number of hits. In conjunction with the model developers, we concluded that the model overestimates the seed particle population as well as the maximum energy that particles can achieve as they are energized, and subsequently made changes to the model to decrease the amount of false alarms.

ZEUS+iPATH	SEPVAL		Scoreboard	
All Clear	>10 MeV	>100 MeV	>10 MeV	>100 MeV
	($N = 60$)	($N = 60$)	($N = 815$)	($N = 867$)
Percent Correct	0.75	0.77	0.92	0.93
Hit Rate	0.65	0.53	0.59	0.67
False Alarm Rate	0.14	0.16	0.063	0.062
False Alarm Ratio	0.17	0.47	0.69	0.87
Bias	0.77	1.0	1.9	5.1
Threat Score	0.57	0.36	0.26	0.12
HSS	0.50	0.37	0.37	0.20
TSS	0.51	0.37	0.53	0.61

Table 6.54: ZEUS+iPATH All Clear Metrics.

To identify the onset peak for iPATH we apply the same algorithm in FetchSEP used to prepare the observed time profiles to the predicted time profiles. This means that the onset peak metrics are calculated only for the scenarios in which iPATH predicts an SEP event above threshold (so the onset peak algorithm can be used) *and* there is an observed SEP event matched to that prediction, i.e., only the hits. It should be noted that the same will be true for SEPMOD, which was also processed using FetchSEP, but that it is different from the SEPSTER and SEPSTER2D analysis which includes every forecast associated with observed SEP events (the hits and misses) for both onset peak and max flux.



(a) >10 MeV (b) >100 MeV

Figure 6.30: Histogram of ZEUS+iPATH onset peak log errors for SEPVAL (blue) the SEP Scoreboards (orange) in the >10 MeV (left) and >100 MeV (right) energy channels.

Table 6.55 shows the metrics for the onset peak flux. For both datasets and both energies, $\geq 50\%$ of the onset peak predictions are within an order of magnitude of the observed peak, but significantly less are within a factor of 2 ($\leq 25\%$). The distributions of the onset peak log error is shown in Figure 6.30. For both energies in the SEPVAL dataset, the distributions are centered near 0 (both slight overpredictions) but both have secondary peaks in the distribution overpredicting by at least an order of magnitude. The distribution for the >100 MeV channel on the Scoreboard is similar to the SEPVAL distribution but without the second peak overpredicting. For the >10 MeV Scoreboard distribution, there is a bimodal distribution at -1 and 1 log error, so under- and over-predicting by an order of magnitude respectively.

ZEUS+iPATH Onset Peak Flux	SEPVAL		Scoreboard	
	>10 MeV ($N=20$)	>100 MeV ($N=8$)	>10 MeV ($N=21$)	>100 MeV ($N=7$)
Percent within a factor of 10	75	50	52	86
Percent within a factor of 2	25	13	4.8	14
Median Log Error	0.33	1.1	0.57	0.35
Median Absolute Log Error	0.48	1.1	0.91	0.49
Spearman Correlation Coefficient	0.50	-0.19	-0.24	0.14

Table 6.55: ZEUS+iPATH Onset Peak Flux Metrics.

The observed maximum flux value can be associated with the arrival of a CME at Earth causing an ESP enhancement, with transport or connectivity effects that cause "wiggles" that generate a high value at a random time, or it could be the

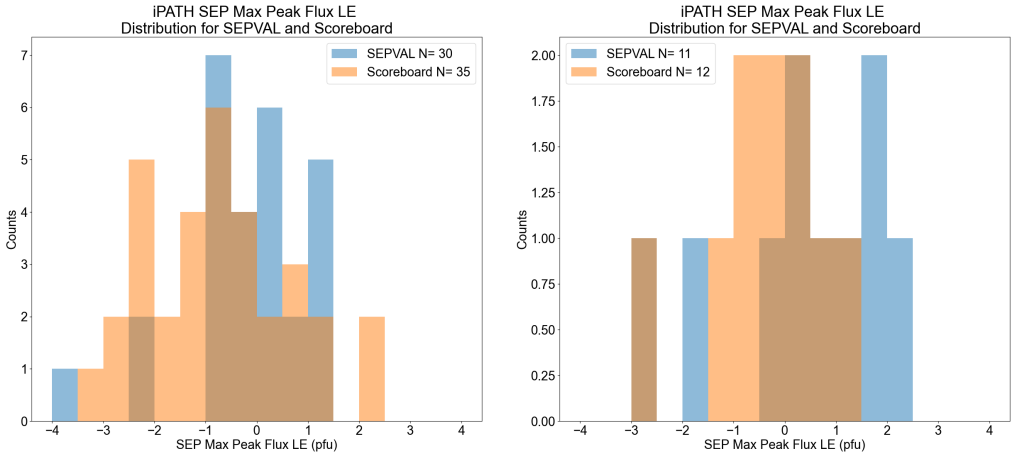
same as the onset peak if the initial peak happened to achieve the maximum flux during the event. The predicted max flux is taken as the maximum value in the predicted time profile. FetchSEP will always extract the maximum flux value in any predicted time profile whether it exceeds operational thresholds or not. This means that the metrics in Table 6.56 are calculated from forecast-observation pairs for all observed SEP events, including cases where the predicted time profile didn't cross threshold and was a miss. For this reason, more predictions (larger N) are included in these metrics compared to onset peak, which only included the subset of forecasts that were hits. Due to these caveats, it is expected that the max flux metrics are worse than those for onset peak. The distributions seen in Figure 6.31 shows a very wide range of log errors with underpredictions by up to -4 orders of magnitude and overpredictions of more than 2 orders of magnitude. The Scoreboard distribution for >10 MeV is fairly symmetrical around -1. The SEPVAL distribution has a peak at the same value, indicating that the model tends to underpredict the max flux. The Spearman correlation coefficient shows low levels of correlation (~ 0.250) between observations and predictions; the model is a poor predictor of event-to-event variability. The >100 MeV distributions both have peaks at 0, but the Scoreboard distribution favors underprediction whereas SEPVAL favors overprediction. Due to these opposite results, as well as the other metrics between the datasets being so different from each other, no general conclusions can be drawn, besides iPATH having little to no skill in capturing the max flux of this energy channel.

ZEUS+iPATH	SEPVAL		Scoreboard	
Max Flux	>10 MeV ($N=31$)	>100 MeV ($N=15$)	>10 MeV ($N=37$)	>100 MeV ($N=12$)
Percent within a factor of 10	63	36	43	58
Percent within a factor of 2	27	27	6	8
Median Log Error	-0.22	0.26	-0.91	-0.60
Median Absolute Log Error	0.77	1.6	1.1	0.79
Spearman Correlation Coefficient	0.25	-0.009	0.26	0.57

Table 6.56: ZEUS+iPATH Max Peak Flux Metrics.

Since iPATH is a physics-based model, requiring the use of an MHD simulation for each prediction, there is an expectation that it will not provide advance warning to the start of an SEP event (especially since it requires human-in-the-loop CME parameters). Therefore the negative advance warning to SEP Start Time in Table 6.57 is expected. However, there is some advance warning of 2.5 hours provided to the onset peak and 4.1 hours to the max flux time for the >10 MeV channel. For the >100 MeV channel, there is typically no advance warning provided ahead of either peak.

One question that SRAG wants to have the answer to in operations is "how long will an SEP event last?" Most of the models we use do not have the capability to give an answer to this question. iPATH is one of the few models on the Scoreboard that attempts to answer this. Table 6.58 shows some simple metrics to show the error in the duration between the model and observations, where a negative (positive)



(a) >10 MeV (b) >100 MeV

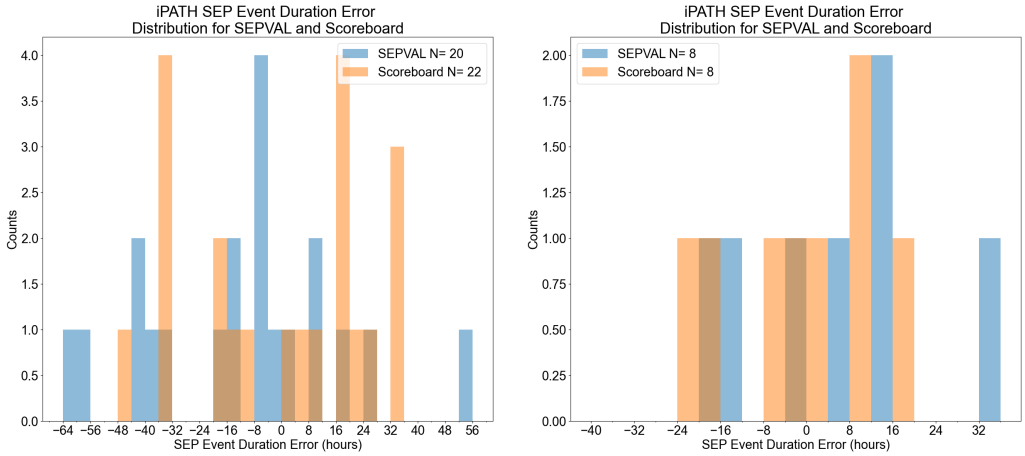
Figure 6.31: Histogram of ZEUS+iPATH max flux log errors for SEPVAL (blue) the SEP Scoreboards (orange) in the >10 MeV (left) and >100 MeV (right) energy channels.

ZEUS+iPATH Advance Warning Time (AWT)	Scoreboard > 10 MeV ($N = 11$)	Scoreboard > 100 MeV ($N = 3$)
AWT to Observed SEP Start Time		
Median	−2.9 hr	−10.0 hr
Worst	−17.3 hr	−17.3 hr
Best	8.6 hr	−2.3 hr
AWT to Observed SEP Onset Peak Time		
Median	2.5 hr	−2.1 hr
Worst	−13.2 hr	−5.7 hr
Best	12.0 hr	1.7
AWT to Observed SEP Max Flux Time		
Median	4.1 hr	−5.7 hr
Worst	−13.2 hr	−12.2 hr
Best	25.5 hr	1.7 hr

Table 6.57: Advance Warning Time for ZEUS+iPATH on the SEP Scoreboard for >10 and >100 MeV.

ZEUS+iPATH Duration	SEPVAL >10 MeV ($N = 20$)	SEPVAL >100 MeV ($N = 8$)	Scoreboard >10 MeV ($N = 22$)	Scoreboard >100 MeV ($N = 8$)
Mean Error	−11.6	−1.4	−3.1	−0.9
Median Error	−6.7	−1.6	5.0	−0.5

Table 6.58: ZEUS+iPATH SEP Event Duration Error (hours)



(a) >10 MeV

(b) >100 MeV

Figure 6.32: Histograms of duration errors for ZEUS+iPATH for the two datasets. Blue is for the SEPVAL dataset and orange is the SEP Scoreboard. Left is >10 MeV, Right is >100 MeV.

sign indicates that the modeled duration is too short (long). For the >10 MeV channel, the mean errors are −11.6 and −3.1 hours for SEPVAL and Scoreboards respectively. The values for >100 MeV are better with means of −1.4 and −0.9 hours respectively. Looking solely at means, however, can be misleading, as the histograms in Figure 6.32 show that the duration errors span multiple days for >10 MeV and ∼ 1 day for the >100 MeV. Due to the spread in these distributions, we can say that iPATH has no skill in capturing the duration of SEP events.

To summarize the performance of ZEUS+iPATH in the SEPVAL challenge and in real time on the SEP Scoreboards:

- All Clear >10 MeV: ZEUS+iPATH shows consistent performance across the two datasets for Hit Rate, percent correct and TSS indicating that iPATH has some skill in determining All Clear, however the False Alarm Ratio is high for the Scoreboards.

- All Clear >100 MeV: Percent correct and Hit Rate are consistent across the datasets, but low Threat Scores and high False Alarm Ratios show that ZEUS+iPATH overpredicts the occurrence of events for >100 MeV.

- Peak Fluxes >10 MeV: For both onset and max peak, most ZEUS+iPATH forecasts ($\geq$ 50%) are within a factor of 10, but significantly less are within a factor of 2. Errors extend beyond 2 orders of magnitude. This shows ZEUS+iPATH has low reliability in determining peak flux. iPATH shows little to no skill in capturing the event-to-event variability of the peak flux values as most correlation coefficients are ≤ 0.25.

- Peak Fluxes >100 MeV: ZEUS+iPATH shows inconsistent performance across the datasets, but marginally better performance for the onset peak, with $\geq$ 50% within a factor of 10. This version of iPATH has no skill in determining the peak of ESPE events, but targeted updates have been made to improve the performance for this energy channel and will be available in the future.

- AWT: ZEUS+iPATH provides no advance warning prior to the start of an SEP event due to the delays in obtaining CME parameters, but does provide advance warning to the onset peak and max flux for the >10 MeV channel. No advance warning is provided to start or peak of >100 MeV events.

6.10 ENLIL+SEPMOD

SEPMOD is a physics-based test particle code for modeling SEP events (Luhmann et al., 2007, 2010). At a fundamental level, it assumes that particles are continuously accelerated and subsequently escape from CME shocks as long as the shock is magnetically connected to the observer as it traverses the inner heliosphere. At each timestep, a sample of test particles is released from the ICME shock with a post-accelerated energy spectrum and are followed until they either arrive at Earth (or other spacecraft) or escape the simulation boundaries. In order to determine the post-acceleration injection spectrum as well as the location of the CME and its magnetic connection to an observer, SEPMOD must be coupled to an MHD solar wind model. Wang-Sheeley-Arge (WSA)-ENLIL provides the solar wind soluation for both SEPVAL and the SEP Scoreboards. WSA-ENLIL generates the solar wind in 3 dimensions throughout the inner heliosphere. The WSA coronal model calculates open and closed field lines on the solar surface then pushes them upwards through the corona. At the inner boundary of ENLIL (21.5 R_{Sun}) any open field lines are allowed to propagate outwards into the heliosphere, creating a solar wind background into which a cone CME can be inserted and propagated to determine shock strength, arrival times, and magnetic connection with time. SEPMOD releases particles from the shock, adjusting the input energy spectrum according to the shock parameters, and tracks test particles along the field line connecting the CME shock to an observer (in our case Earth). The particles arriving at the observer are counted throughout the full simulation time (nominally 7 days), generating flux time profiles for multiple energy channels of interest: >10 MeV, >30 MeV, >50 MeV, > 60 MeV, >100 MeV, >300 MeV, >500 MeV, >750 MeV. The version of SEPMOD used in both the SEPVAL challenge and on the SEP Scoreboard includes the optional feature to inject a secondary soft power law spectrum associated with the arrival of a CME shock as part of the ESP phase.

To interpret the validation results from SEPMOD, we need to first address the known caveats of SEPMOD, as well as any caveats for WSA-ENLIL. Firstly, SEPMOD assumes particles only arrive from the field line that is magnetically connected at each timestep (this field line can change between timesteps but the particles from the previous timestep will still arrive), which means the perpendicular diffusion of particles is neglected entirely. This limitation causes the model to miss some SEP events when there's no modeled magnetic connection between the CME and the observer, e.g., some Eastern or behind-the-limb CMEs. Additionally, for real-time runs on the SEP Scoreboards, SEPMOD applies a cut to avoid running for slow CMEs (below < 450 km/s) which typically aren't associated with SEP events. WSA-ENLIL also has multiple caveats to consider. Firstly, the inner boundary where CMEs are initially injected is at 21.5 R_{Sun}, where it is assumed all open (closed) field lines will remain open (closed) as they propagate outwards. It is theorized that energetic particles are accelerated much closer to the Sun where the initial strong shock formation occurs, far below ENLIL's inner boundary, implying that initial onset of SEP events cannot be simulated properly in the coupled models. Additionally, WSA-ENLIL uses a cone model to simulate CMEs, which does not contain a

ENLIL+SEPMOD		
Characteristic	> 10 MeV	> 100 MeV
First Forecast	2021/05/07	2021/05/07
Last Forecast	2024/12/25	2024/12/25
N Days	1350	1350
N Forecast Days	982	982
N SEP Days	29	8
Forecast Cadence	Triggered	Triggered
Prediction Window	7 days	7 days
N forecasts	2045	2045
N matched w/events	43	12
Imbalance (raw)	46.56	169.42
Imbalance (days)	32.86	121.75

Table 6.59: ENLIL+SEPMOD validation characteristics table.

magnetic flux rope structure in the CME and therefore lacks a magnetic component that can interact with the solar wind, simplifying the CME's interactions and evolution. This simplification affects the injected spectrum of particles in SEPMOD, where stronger shocks lead to harder spectra and more high energy particles, as well as the dynamics of the connected field line affecting the transport of the test particles.

SEPMOD produces a time profile for each input CME, which is processed with FetchSEP to extract All Clear, peak intensity, timing information, and more. The All Clear binary is determined by checking whether the time profile crosses operational thresholds (>10 MeV at 10 pfu and >100 MeV at 1 pfu). Contingency tables are reported in Table 6.60 and metrics are provided in Table 6.61. SEPMOD shows some skill in discriminating between SEP events and non-event periods. For SEPVAL, SEPMOD's percent correct corresponds to the median value of the SEPVAL models. The >10 MeV Hit Rate is slightly lower than median, but the False Alarm Rate is better than median. This points to a tendency to underpredict the >10 MeV fluxes. The Hit Rate and False Alarm Rate are very small for >100 MeV, indicating an even stronger tendency towards underprediction. Indeed, Bias values < 1 support this conclusion. The HSS and TSS for >10 MeV of $\sim$0.46 are equivalent to the median skill of the SEPVAL models, however the >100 MeV scores < 0.3 are worse than median performance. For the SEP Scoreboards, the percent correct increases due to the large number of correct negatives, however the Hit Rate drops very significantly with 26% for >10 MeV and no hits (0%) for >100 MeV. The reason for the significant drop in performance across datasets is unclear, but we note that the SEPVAL dataset has SEP events associated with many more wide CMEs compared to the SEP Scoreboard events, as shown in Figure 3.5. These wider CMEs have a better chance of making a magnetic connection to the observer and increasing the number of hits. This implies that SEPMOD is highly sensitive to magnetic connectivity predictions and performs better for wider CMEs with source eruptions at better connected solar longitudes, also noted in Palmerio et al. (2022);

SEPVAL

> 10 MeV

	Observed Yes	No	Sum
Pred. Yes	19	4	23
Pred. No	13	26	39
Sum	32	30	62

> 100 MeV

	Observed Yes	No	Sum
Pred. Yes	6	4	10
Pred. No	11	38	49
Sum	17	42	59

Scoreboard

> 10 MeV

	Observed Yes	No	Sum
Pred. Yes	11	95	106
Pred. No	32	1884	1916
Sum	43	1979	2022

> 100 MeV

	Observed Yes	No	Sum
Pred. Yes	0	0	0
Pred. No	12	2031	2043
Sum	12	2031	2043

Table 6.60: ENLIL+SEPMOD contingency tables.

ENLIL+SEPMOD All Clear	SEPVAL > 10 MeV ($N = 62$)	SEPVAL > 100 MeV ($N = 59$)	Scoreboard > 10 MeV ($N = 2022$)	Scoreboard > 100 MeV ($N = 2043$)
Percent Correct	0.73	0.75	0.94	0.994
Hit Rate	0.59	0.35	0.26	0.0
False Alarm Rate	0.13	0.10	0.05	0.0
False Alarm Ratio	0.17	0.40	0.90	Undefined
Bias	0.72	0.59	2.47	0.0
Threat Score	0.53	0.29	0.08	0.0
HSS	0.46	0.29	0.12	0.0
TSS	0.46	0.26	0.21	0.0

Table 6.61: ENLIL+SEPMOD All Clear Metrics.

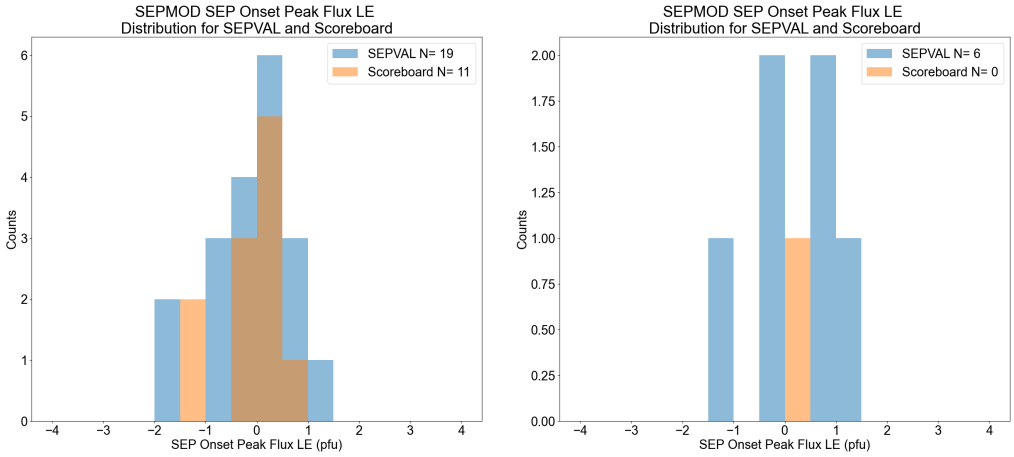
Whitman et al. (2023).

As SEPMOD produces a time profile, it can be used to make predictions for both the onset peak flux Table 6.62 and maximum peak flux Table 6.63. As with iPATH, SEPMOD time profiles are processed with FetchSEP. FetchSEP will only identify an onset peak when SEPMOD predicts that there will be a threshold crossing, however max flux will be extracted for every predicted time profile. This results in onset peak metrics composed of scenarios when an SEP event is predicted and observed, i.e., hits. The max flux metrics are calculated from all predictions associated with observed SEP events, both hits and misses. As noted previously, this is different than SEPSTER and SEPSTER2D for which both onset peak and max flux were calculated from every forecast associated with observed SEP events (hits and misses).

Starting with onset peak, SEPMOD shows some skill across all metrics in both energy channels for the SEPVAL dataset, with 84% of >10 MeV predictions and 67% of >100 MeV predictions within a factor of 10 and 1/3 within a factor of 2 for both energies. The histograms in Figure 6.33 shows that for the >10 MeV channel both for the Scoreboard and SEPVAL, the distributions are centered near 0

ENLIL+SEPMOD Onset Peak Flux	SEPVAL		Scoreboard	
	>10 MeV ($N=19$)	>100 MeV ($N=6$)	>10 MeV ($N=11$)	>100 MeV ($N=0$)
Percent within a factor of 10	84	67	82	-
Percent within a factor of 2	32	33	46	-
Median Log Error	0.15	0.36	0.16	-
Median Absolute Log Error	0.47	0.86	0.36	-
Spearman Correlation Coefficient	0.41	0.14	-0.14	-

Table 6.62: ENLIL+SEPMOD Onset Peak Flux Metrics.



(a) >10 MeV

(b) >100 MeV

Figure 6.33: Histogram of ENLIL+SEPMOD onset peak log errors for SEPVAL (blue) the SEP Scoreboards (orange) in the >10 MeV (left) and >100 MeV (right) energy channels.

(median log error of 0.15) with a longer tail in the negative direction. For the >100 MeV channel, since there are no hits for the SEP Scoreboards, the histogram only shows the SEPVAL distribution, which is centered at 0 (median log error of 0.16). The Spearman correlation coefficient is used to assess how the model performs with the event-to-event variability, here the >10 MeV metric shows a medium level of correlation between predictions and observations, but little to no correlation is seen for the >100 MeV metric.

The maximum peak flux during an SEP event can sometimes be associated with a shock arrival at the observer leading to an ESP enhancement of energetic particles accelerated locally at the shock. SEPMOD can attempt to model this ESP phase by injecting a secondary soft spectrum of particles when a CME is close to the observer, sometimes resulting in a maximum flux value during the modeled ESP. For maximum peak flux, SEPMOD performs worse than it does for onset peak, although this is to be expected due to the inclusion of misses in this metric. The >100 MeV predictions show a large underestimate of over an order of magnitude for

ENLIL+SEPMOD Max Flux	SEPVAL		Scoreboard	
	> 10 MeV ($N = 32$)	> 100 MeV ($N = 17$)	> 10 MeV ($N = 43$)	> 100 MeV ($N = 12$)
Percent within a factor of 10	63	36	68	33
Percent within a factor of 2	26	14	26	0
Median Log Error	-0.60	-1.1	-0.59	-1.4
Median Absolute Log Error	0.79	1.2	0.73	1.4
Spearman Correlation Coefficient	0.32	0.05	-0.17	0.53

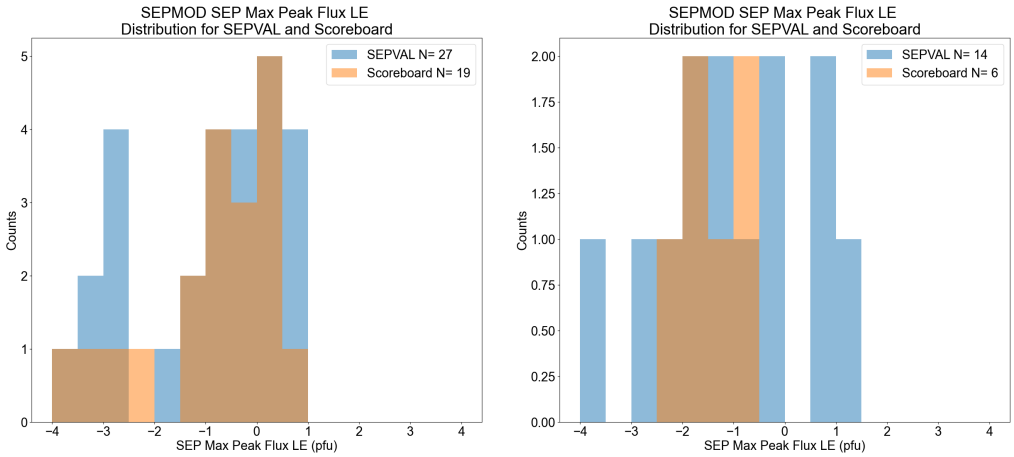
Table 6.63: ENLIL+SEPMOD Max Flux Metrics.

both SEPVAL and the Scoreboards. For the >10 MeV channel, for both datasets, approximately 2/3 of the predicted max flux is within a factor of 10, with 1/4 being within a factor of 2. Median log error is -0.602 (-0.59 for the Scoreboard) which is about a factor of ~ 4 with a bias towards underprediction, and median absolute log error is 0.79 (0.73) is a factor of ~ 6. The Spearman coefficient is 0.32 which shows only a low level of correlation for SEPVAL and -0.17 (no positive correlation) for the Scoreboard, meaning SEPMOD does not capture the event-to-event variability of the maximum peak flux. The histogram distributions in Figure 6.34 for the >10 MeV protons are centered near zero for both datasets, but, for SEPVAL, there is an additional peak in the distribution at -3. Sometimes SEPMOD produces all 0 time profiles, indicating that a magnetic connection was never made between the observer and the shock. Because the peak flux scores are calculated in log space, these all-zero predictions are excluded from the peak flux calculations and figures. They are, however, penalized in the All Clear metrics.

The correlation plot in Figure 6.35 shows the log error as a function of predicted flux. The plot demonstrates that the largest errors are primarily due to significant underpredictions of the observed flux values. This may be partially due to the lack of perpendicular diffusion in the model and the strict requirement of a direct magnetic connection to the shock for particles to arrive at the observer.

As a physics-based model, there is an expectation from SEPMOD that its computation time exceeds that of the empirical or machine learning models, which limits the warning the model can provide. Table 6.64 demonstrates that ENLIL+SEPMOD generally does not provide advance warning to the start or onset peak of SEP events, issuing forecasts a median of 3.5 and 1.5 hours afterward, respectively. However, the model is able to issue predictions prior to the maximum flux time with a median advance warning of 0.5 hours.

Despite the lack of warning that SEPMOD provides, we can still use this model to help provide situational awareness, and answer other questions an operator may have during the course of an SEP event, like "how long will an SEP event last?" The duration metrics for SEPMOD are shown in Table 6.65 and the histograms in Figure 6.36 shows the error distribution. For >10 MeV, SEPMOD's predicted event duration is, on average, $\sim$30 hours too short for SEPVAL and $\sim$13 hours too short for the Scoreboard. As noted above, this is likely due to the strict requirement that particles follow field lines from the shock to the observer without accounting



(a) >10 MeV (b) >100 MeV

Figure 6.34: Histogram of ENLIL+SEPMOD max flux log errors for SEPVAL (blue) the SEP Scoreboards (orange) in the >10 MeV (left) and >100 MeV (right) energy channels.

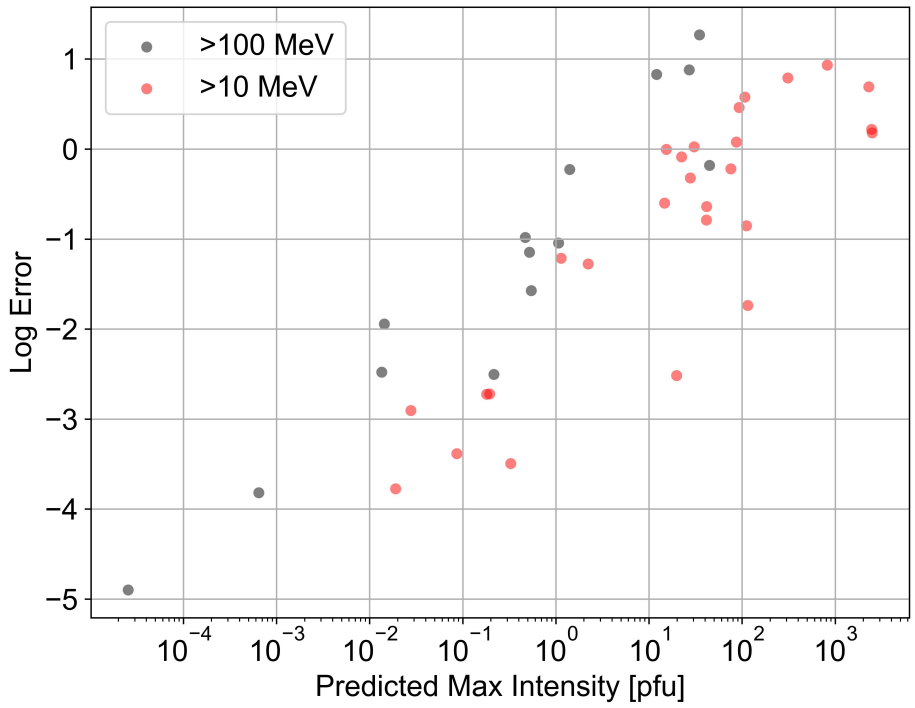


Figure 6.35: ENLIL+SEPMOD max flux prediction versus associated log error for the SEPVAL dataset.

ENLIL+SEPMOD Advance Warning Time (AWT)	Scoreboard > 10 MeV ($N = 8$)	Scoreboard > 100 MeV ($N = 0$)
AWT to Observed SEP Start Time		
Median	−3.5 hr	-
Worst	−148 hr	-
Best	14.5 hr	-
AWT to Observed SEP Onset Peak Time		
Median	−1.5 hr	-
Worst	−146.9 hr	-
Best	23.4 hr	-
AWT to Observed SEP Max Flux Time		
Median	0.5 hr	-
Worst	−144.3 hr	-
Best	42.5 hr	-

Table 6.64: Advance Warning Time for ENLIL+SEPMOD on the SEP Scoreboards for >10 and >100 MeV.

ENLIL+SEPMOD Duration	SEPVAL > 10 MeV ($N = 20$)	SEPVAL > 100 MeV ($N = 6$)	Scoreboard >10 MeV ($N = 11$)	Scoreboard >100 MeV ($N = -$)
Mean Error	−32.6	−13.2	−12.9	-
Median Error	−30.7	−5.5	−0.6	-

Table 6.65: ENLIL+SEPMOD Duration Metrics

for the arrival of particles due to diffusion across field lines. Generating ensemble predictions over a range of CME parameters (in work) may improve performance.

To summarize:

- All Clear >10 MeV: ENLIL+SEPMOD shows medium skill in differentiating between non-event and event periods in the SEPVAL dataset, with an HSS of 0.46 that is equivalent to the median skill of the SEPVAL models. With a Hit Rate of only 26% for the Scoreboards, SEPMOD has no skill in predicting SEP event occurrence in real time. The high False Alarm Ratio indicates that, if the model predicts there will be an event, 90% of the time it is a false alarm.

- All Clear >100 MeV: For SEPVAL, ENLIL+SEPMOD has low skill in identifying the conditions that cause >100 MeV SEP events. The low Hit Rate of 35% and high False Alarm Ratio of 40% are worse than the median performance of participating models. On the Scoreboards, ENLIL+SEPMOD did not predict a single >100 MeV event threshold crossing during the Scoreboard era.

- Peak Fluxes >10 MeV: ENLIL+SEPMOD shows some skill in predicting the peak flux of an SPE event. The model performs well for onset peak magnitude,

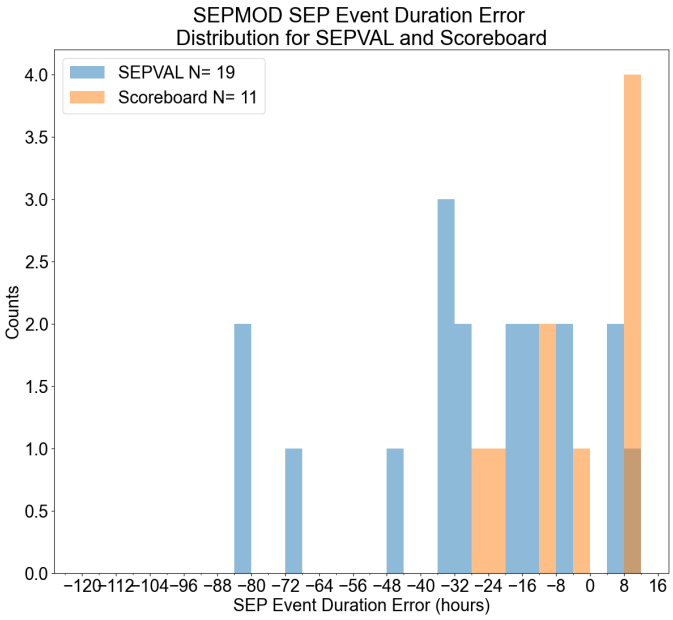


Figure 6.36: ENLIL+SEPMOD duration error distributions for >10 MeV SEP events. Blue is SEPVAL and Orange is Scoreboard.

with most predictions (∼80%) in both datasets within a factor of 10 and 1/3 within a factor of 2. SEPVAL predictions have medium correlation with observations. The Scoreboard results, however, show weak negative correlation indicating that event-to-event variability is not well-predicted in real time. The max flux predictions show overall lower scores but similar trends.

- Peak Fluxes >100 MeV: ENLIL+SEPMOD has slightly worse performance than the >10 MeV channel for onset peak metrics, but significant decrease in performance for max peak metrics, with median predictions being a order of magnitude too low with the inclusion of 'misses' in the metric. Event-to-event variability measured with Spearman Correlation shows low to no skill except for max peak on the Scoreboards which has medium correlation despite all forecasts being underpredictions.

- AWT: ENLIL+SEPMOD provides no advance warning for SEP start time nor onset peak time, but some minor (∼30 min) warning to maximum flux time.

6.11 UMASEP

The UMASEP models are an empirical suite of tools comprised of multiple independent modules predicting different particle energy ranges: UMASEP-10 (Núñez, 2011) predicts >10 MeV SEP events; UMASEP-100 (Núñez, 2015) predicts >100 MeV events; and UMASEP-500 (Núñez et al., 2017) predicts the occurrence of a Ground-Level Event (GLE). UMASEP-500 was developed in collaboration with the European Unions's HESPERIA project, which also includes the REleASE code (Section 6.12). Two additional modules, UMASEP-30 and UMASEP-50, have recently been developed to forecast >30 and >50 MeV events. All model versions use similar underlying algorithms; however, the implementation differs.

The forecast functionality of UMASEP-10 is divided into the Well-Connected (WC), Poorly-Connected (PC), and University of Málaga predictor from Solar Data (UMASOD) modules. The WC module assesses the relationship between the time series GOES Soft X-Ray (SXR) (X-ray emissions) and an increase in GOES differential proton flux to determine if particles have escaped along the IMF field lines that are well-connected to the observer. The PC module aims to identify precursors of events with a poor magnetic connection between the observer and the source, using a regression model that checks whether the differential proton flux behavior is similar to the beginning phase of previous historically poorly connected SEP events. The UMASOD model incorporates radio data (Núñez and Paul-Pena, 2020; Zucca et al., 2017) to enhance the model's ability to provide advance warning of an SEP event. The preliminary warnings from the three modules are passed to the Analysis and Inference module with preference assigned to the WC module; if this preliminary result is accepted, an alert is issued.

UMASEP-100 performs in a similar manner to the WC module, but uses a derived instead of real-value time series. The updated algorithm performs a bit-based transformation of the X-ray flux and the GOES differential proton channels P6 to P11, with a focus on strong positive derivatives (1 vs 0). The newer UMASEP-30 and UMASEP-50 modules were developed based on a similar theory as the original UMASEP-10 and UMASEP-100.

UMASEP-500 is very similar to the UMASEP-100 method with the primary goal of forecasting whether a GLE will occur or not. Due to the very small number of >500 MeV events in Solar Cycle 25, this version of UMASEP is not validated in this report.

Table 6.66 lists the characteristics of UMASEP's forecasting statistics on the SEP Scoreboards. UMASEP issues a new forecast every 3 - 5 minutes, producing millions of forecasts across its multiple models. UMASEP makes predictions for All Clear, threshold crossing time, and maximum flux in the prediction window. UMASEP's prediction window is designed to encompass the first few hours of an SEP event, so the flux prediction is most similar to observed onset peak. Here we focus on All Clear and onset peak results for UMASEP-10 and UMASEP-100.

Tables 6.67 and 6.68 report the UMASEP-10 and UMASEP-100 contingency tables for individual forecasts and forecasts deoverlapped into daily periods. The deoverlapping method applied is described in detail in Section 3.5. The resulting

	UMASEP	
Characteristic	>10 MeV	>100 MeV
First Forecast	2020/03/13	2020/03/13
Last Forecast	2024/12/31	2024/12/31
N Days	1755	1755
N Forecast Days	1659	1669
N SEP Days	35	8
Forecast Cadence	3 min	3 min
Prediction Window	7 - 24 hrs	3 hrs
N forecasts	810,486	841,771
N matched w/events	2269	226
Imbalance (raw)	356	3724
Imbalance (days)	46.4	207.6

Table 6.66: UMASEP-10 and UMASEP-100 validation characteristics.

All Clear metrics are shown in Tables 6.69 and 6.70 for both cases. Immediately, the positive effect of deoverlapping on the skill scores is evident. The Hit Rate is increased, the False Alarm Ratio is reduced and the Bias, Threat Score, HSS, and TSS are improved. An explanation is found in Figure 6.37. An observed >10 MeV event (red line) on the Intensity SEP Scoreboard is overlaid with UMASEP-10 forecasts, shown as red triangles with associated prediction windows plotted as shaded bars. In this case, all forecasts issued predict the same peak flux value, causing them to overlap. Here, and in many cases on the SEP Scoreboard, UMASEP correctly forecasts that an SEP event will occur. It issues forecasts throughout the event and then continues to issue Not Clear and peak flux forecasts even though the observed event has already dropped below threshold. These "trailing" false alarms are issued every 3 - 5 minutes, racking up hundreds of false alarms in a short period of time. Because of the tendency for UMASEP's false alarms to follow a pattern, deoverlapping reduces the ratio of false alarms to hits and benefits the skill scores. As an example, deoverlapping converts the 107 false alarms on May 30 and 227 false alarms on May 31 to one false alarm for each 24 hour period. Not all of UMASEP's false alarms follow this pattern, but for >10 MeV, "trailing" false alarms occur for up to 10 SEP events. If this undesirable behavior was surpressed, the False Alarm Rate and Ratio could be significantly improved, particularly in the "raw" scores.

With the deoverlapped approach, UMASEP-10 and -100 demonstrate very high skill, producing at least one hit for for 69% (24/35) of >10 MeV events and 75% (6/8) of >100 MeV events. The False Alarm Rates are low with false alarms on only 3% of days for >10 MeV and 0.9% of days for >100 MeV. Due to the extreme imbalance of SEP climatology, the False Alarm Ratios are still high at 63% and 67%, respectively, nonetheless, the skill scores are very good. With Threat Score = 0.32 (0.30), HSS = 0.46 (0.46), and TSS = 0.65 (0.74) for UMASEP-10 (-100), UMASEP achieves the highest All Clear scores out of all the models on the SEP Scoreboards, with the possible exception of HESPERIA REleASE (see Section 6.12).

Similar high skill was seen for SEPVAL. Table 6.71 reports the deoverlapped

>10 MeV

Scoreboard Individual Forecasts

	Observed		
	Yes	No	Sum
Pred. Yes	1138	8711	9849
Pred. No	1131	797,871	799,002
Sum	2269	806,582	808,851

Scoreboard Deoverlapped

	Observed		
	Yes	No	Sum
Pred. Yes	24	41	65
Pred. No	11	1285	1296
Sum	35	1326	1361

Table 6.67: Contingency tables for UMASEP-10 for the total number of individual forecasts submitted to the SEP Scoreboards (left) and the deoverlapped results for 24 hour periods (right).

>100 MeV

Scoreboard Individual Forecasts

	Observed		
	Yes	No	Sum
Pred. Yes	44	1974	2018
Pred. No	182	839,411	839,593
Sum	226	841,385	841,611

Scoreboard Deoverlapped

	Observed		
	Yes	No	Sum
Pred. Yes	6	12	18
Pred. No	2	1332	1334
Sum	8	1344	1352

Table 6.68: Contingency tables for UMASEP-100 for the total number of individual forecasts submitted to the SEP Scoreboards (left) and the deoverlapped results for 24 hour periods (right).

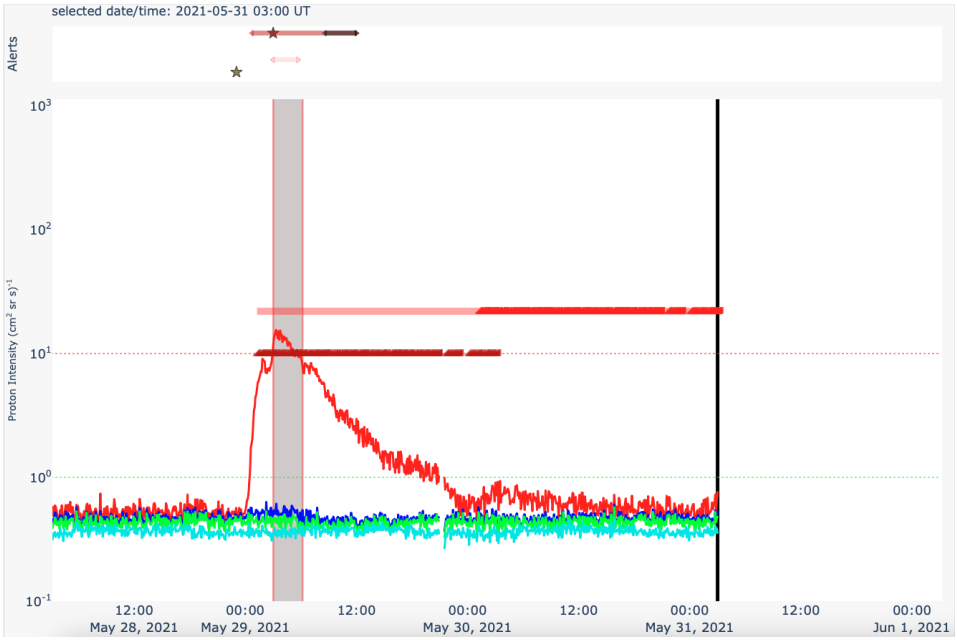


Figure 6.37: UMASEP-10 has a tendency to continue forecasting not clear after an observed SEP event has ended, producing numerous false alarms.

UMASEP-10 All Clear SEP Scoreboard	Individual Forecasts >10 MeV ($N = 808,851$)	Deoverlapped >10 MeV ($N = 1361$)
Percent Correct	0.988	0.96
Hit Rate	0.50	0.69
False Alarm Rate	0.011	0.03
False Alarm Ratio	0.88	0.63
Bias	4.34	1.86
Threat Score	0.10	0.32
HSS	0.18	0.46
TSS	0.49	0.65

Table 6.69: UMASEP-10 All Clear metrics for individual and deoverlapped forecasts on the SEP Scoreboards.

UMASEP-100 All Clear SEP Scoreboard	Individual Forecasts >100 MeV ($N = 841,611$)	Deoverlapped >100 MeV ($N = 1352$)
Percent Correct	0.997	0.99
Hit Rate	0.19	0.75
False Alarm Rate	0.002	0.009
False Alarm Ratio	0.98	0.67
Bias	8.92	2.25
Threat Score	0.02	0.3
HSS	0.039	0.46
TSS	0.19	0.74

Table 6.70: UMASEP-100 All Clear metrics for individual and deoverlapped forecasts on the SEP Scoreboards.

SEPVAL UMASEP-10

	Observed		
	Yes	No	Sum
Pred. Yes	29	1	30
Pred. No	2	29	31
Sum	31	30	61

SEPVAL UMASEP-100

	Observed		
	Yes	No	Sum
Pred. Yes	16	4	20
Pred. No	1	41	42
Sum	57	45	62

Table 6.71: Contingency tables for UMASEP-10 and UMASEP-100 for deoverlapped forecasts for each SEPVAL challenge period.

SEPVAL All Clear	UMASEP-10 >10 MeV ($N = 61$)	UMASEP-100 >100 MeV ($N = 62$)
Percent Correct	0.95	0.92
Hit Rate	0.94	0.94
False Alarm Rate	0.033	0.089
False Alarm Ratio	0.033	0.20
Bias	0.97	1.18
Threat Score	0.91	0.76
HSS	0.90	0.81
TSS	0.90	0.85

Table 6.72: UMASEP-10 and UMASEP-100 All Clear metrics for deoverlapped forecasts for each SEPVAL challenge period.

contingency tables for the SEPVAL time periods. SEPVAL evaluated forecasts for the periods leading up to the observed peak time for each SEP event, so the "trailing" false alarm behavior is not reflected in these metrics. By far, UMASEP had the highest Hit Rate, lowest False Alarm Rate, and highest skill scores of all the participating SEPVAL models, reported in Table 6.72. It should be kept in mind that the model developer submitted forecasts from the most recent version of UMASEP (v3.3) at the time and that the model training likely included many of the SEPVAL challenge periods. Still, UMASEP shows a strong performance for binary All Clear across both SEPVAL and the SEP Scoreboards.

Metrics describing UMASEP's performance for peak flux prediction are reported in Table 6.73. Figures 6.38 and 6.39 show correlation plots for predicted peak versus observed onset peak for UMASEP-10 for SEPVAL (left) and the SEP Scoreboard (right). Figures 6.40 and 6.41 show the same correlation plots for UMASEP-100. Deoverlapping is not applied for the peak flux metrics here and they may include multiple forecasts for the same observed event. SPHINX calculates metrics for various subsets — first forecasted flux, maximum forecasted flux, and mean forecasted flux per event — but these results are not shown.

The correlation plots show that the peak flux predictions are highly correlated with observations in the SEPVAL dataset, but on the SEP Scoreboards, the predic-

UMASEP Onset Peak Flux	SEPVAL		Scoreboard	
	>10 MeV ($N=693$)	>100 MeV ($N=122$)	>10 MeV ($N=1050$)	>100 MeV ($N=42$)
Percent within a factor of 10	1.0	0.93	0.83	0.86
Percent within a factor of 2	0.76	0.81	0.5	0.43
Median Log Error	0.16	−0.07	0.08	−0.21
Median Absolute Log Error	0.37	0.13	0.24	0.21
Spearman Correlation Coefficient	0.84	0.70	0.22	−0.06

Table 6.73: UMASEP Onset Peak Flux Metrics.

tions do not demonstrate this level of skill. On the SEP Scoreboards, UMASEP-10 appears to revert to a default prediction of ≈16 pfu for most events (Figure 6.39). The predictions with higher intensities do fall fairly close to the 1:1 trend line, except for one predicted outlier above 1000 pfu. Similar behavior is seen for the SEP Scoreboard UMASEP-100 predictions (Figure 6.41). Most of the SEP events in Solar Cycle 25 have been small, whereas numerous SEPVAL events exceed 1000 pfu for >10 MeV and above 10 pfu for >100 MeV. Most of the observed >10 MeV events are below 100 pfu and no >100 MeV event has exceeded 10 pfu on the SEP Scoreboard. It is for this reason that UMASEP has managed to achieve fairly high performing metrics in Table 6.73 despite the lack of correlation between the predictions and observations — UMASEP's default prediction of 16 pfu is within an order of magnitude of most of the >10 MeV Scoreboard events. Likewise, since the events are small for >100 MeV, UMASEP-100's predictions are also all within an order of magnitude of the observed onset peaks. Looking at the smaller events, particularly for >100 MeV, similar scatter is seen in UMASEP-100's SEPVAL forecasts.

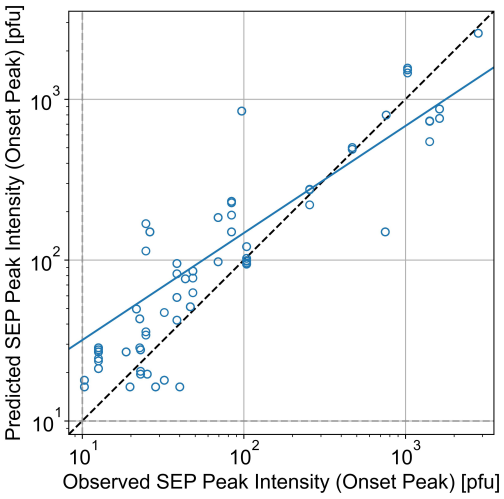


Figure 6.38: UMASEP-10 onset peak correlation for SEPVAL.

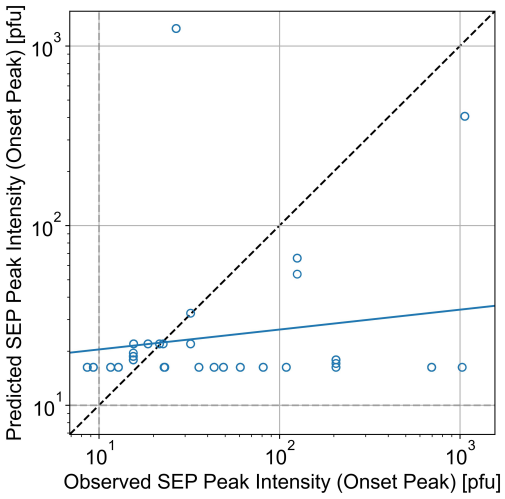


Figure 6.39: UMASEP-10 onset peak correlation for SEP Scoreboard

Both UMASEP-10 and UMASEP-100 provide advance warning ahead of im-

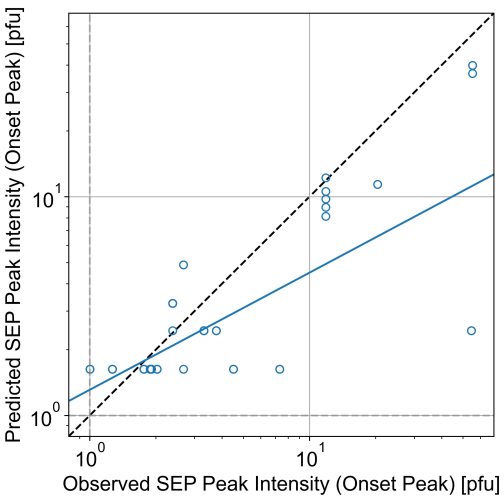


Figure 6.40: UMASEP-100 onset peak correlation for SEPVAL.

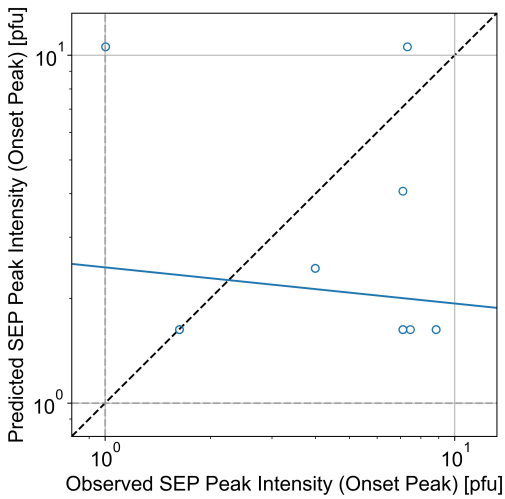


Figure 6.41: UMASEP-100 onset peak correlation for SEP Scoreboard

pending SEP events, as shown in Table 6.74. The median AWT is 44.4 minutes for >10 MeV events and 17.4 minutes for >100 MeV events. Peak forecasts are issued a median of 5.7 and 4.4 hours ahead of observed >10 MeV and >100 MeV onset peak times. AWT for each event is plotted in the box plots in Figure 5.45, which also highlights that UMASEP's advance warning is very similar to that of the SWPC Warning product. UMASEP's ability to issue a forecast minutes to hours ahead of an SEP event, coupled with its relatively high level of skill make it a potentially useful model for operations, particularly if the real-time peak flux predictions can be improved to resemble the performance seen in SEPVAL.

In summary,

- UMASEP-10 and -100 achieve the highest (deoverlapped) skill scores out of all models for both SEPVAL (HSS = 0.90, 0.81) and the SEP Scoreboards (HSS = 0.46, 0.46). Hit Rates are high and vary from 69% - 94%. Real-time False Alarm Ratios are still high at 63% and 67%, due to the extreme imbalance of SEP climatology though the real-time False Alarm Rates are low with false alarms on only 3% of days for >10 MeV and 0.9% of days for >100 MeV.

- UMASEP peak flux predictions are highly correlated with observations in the SEPVAL dataset, but for the SEP Scoreboards, the predictions do not demonstrate the same level of skill. On the SEP Scoreboards, UMASEP-10 often issued a default prediction of ∼16 pfu for most events. Since the events in SC 25 have been small, most forecasts are within an order of magnitude, but not well-correlated, with observed peak fluxes.

- UMASEP provides advance warning ahead of SEP events, even for >100 MeV threshold crossings. The median AWT is 44.4 minutes for >10 MeV events and 17.4 minutes for >100 MeV events.

UMASEP Advance Warning Time (AWT)	Scoreboard > 10 MeV ($N = 23$)	> 100 MeV ($N = 6$)
AWT to Observed SEP Start Time		
Median	44.4 min	17.4 min
Worst	10 sec	3.5 min
Best	18.8 hr	35.4 min
AWT to Observed SEP Onset Peak Time		
Median	5.7 hr	4.4 hr
Worst	1.1 hr	2.7 hr
Best	23.9 hr	4.9 hr

Table 6.74: UMASEP Advance Warning Time
Advance Warning Time for UMASEP on the SEP Scoreboard for >10 and >100 MeV

- UMASEP's ability to issue a forecast minutes to hours ahead of an SEP event, coupled with its relatively high level of skill make it a potentially useful model for operations, particularly if the real-time peak flux predictions can be improved.

6.12 HESPERIA REleASE

The HESPERIA-REleASE model takes advantage of promptly arriving near-relativistic as well as relativistic electrons as an early indicator of a solar energetic particle event. It uses 175 - 315 keV electron measurements from ACE/EPAM and 0.25 - 1 MeV electrons measured by SOHO/Comprehensive Suprathermal and Energetic Particle Analyzer (COSTEP) to predict 15.6 - 39.8 MeV and 28.2 - 50.1 MeV proton fluxes in the 30, 60 and 90 minute future windows. Relativistic electrons ∼1 MeV always arrive at 1 AU prior to the ∼30 - 100 MeV protons (Posner, 2007). For an ideally connected particle event in which particles follow the nominal Parker Spiral of 1.2 AU from the Sun to the Earth at zero pitch angle, 1 MeV electrons would arrive 30.4 minutes earlier than 30 MeV protons, 21.8 minutes earlier than 50 MeV protons, and 13.3 minutes ahead of 100 MeV protons, but would lag behind 300 MeV protons by more than 5 minutes (Posner, 2007). While this method does not provide early warning for the highest energy SEP protons arriving at Earth (>300 MeV), it does provide advance warning for the bulk of SEP events as well as acts as an indicator of magnetic connectivity between the Sun and Earth, i.e., if energetic electrons arrive at Earth, it is clear that particles are able to transport from the active region to the Earth and energetic protons are expected. A unique aspect of this forecast method is inherent in the use of *in situ* particle measurements (i.e., electrons at L1) as opposed to reliance on remote solar observations (e.g., X-rays). HESPERIA REleASE can therefore forecast SEP events associated with solar eruptions behind the western solar limb, an expansion of capabilities beyond those offered by models relying on remote sensing observations only available for the visible solar disc.

HESPERIA REleASE forecasts for protons with energies in the range of 15.8 – 39.8 MeV and 28.2 – 50.1 MeV to the SEP Scoreboards whereas SRAG monitors >10 MeV and >100 MeV GOES protons for operational decisions. The REleASE team chooses not to forecast any of the GOES energetic proton channels due to side-penetration issues identified in Smart and Shea (1999) and the appendix of Posner (2007) and thus limits itself to forecasting <50 MeV protons measured by the cleaner SOHO/COSTEP-EPHIN instrument. However, because SRAG is interested in the GOES channels, the goal in this analysis is to determine whether there is a clear relationship between the differential flux forecasts produced by HESPERIA REleASE and the GOES integral flux channels used in SRAG operations. **The evaluation presented here is not strictly a validation since there is an inherent mismatch in energy channels and thresholds between forecasts and observations, but it is rather a comparison between HESPERIA REleASE's warning conditions for 15.8–39.8 MeV proton flux and GOES >10 MeV threshold crossings.**

HESPERIA REleASE ACE 60-Min and SOHO 60-Min predict the proton flux expected in the near future (∼1 to a few hours). A forecast for 15.8–39.8 MeV proton flux exceeding 0.1 [MeV cm^2 s sr]$^{-1}$ is considered a warning condition. Forecasts for this energy channel and threshold combination were compared to GOES >10 MeV fluxes exceeding 10 pfu to understand if there is a relationship between the two and whether HESPERIA RelEASE provides advance warning with these alert conditions.

	HESPERIA REleASE	
Characteristic	ACE 60-Min	SOHO 60-Min
First Forecast	2020/03/16	2020/03/17
Last Forecast	2024/05/31	2024/05/31
N Days	1538	1537
N Forecast Days	1360	1438
N SEP Days	26	22
Forecast Cadence	5 min	5 min
Prediction Window	20 min - 20 hrs	20 min - 20 hrs
N forecasts	184802	382676
N matched w/events	316	707
Imbalance (raw)	584	540
Imbalance (days)	51.3	64.4

Table 6.75: Forecast characteristics for HESPERIA REleASE ACE 60-Min and SOHO 60-Min on the SEP Scoreboards.

Table 6.75 lists the statistics of the forecasts made to the SEP Scoreboards. The ACE and SOHO detectors both experience data gaps, resulting in different forecast coverage, as shown in the table.

The SEP Scoreboard also includes the HESPERIA REleASE ACE and SOHO 30-Min and 90-Min predictions. Note that the HESPERIA REleASE predictions available on the developers' website[16] require threshold crossings in multiple of the 30-Min, 60-Min, and 90-Min versions in order to predict a Warning condition. This logic was not included in the version of HESPERIA REleASE deployed to the SEP Scoreboard, so the 60-Min version is analyzed as a standalone module.

The HESPERIA REleASE threshold crossings warnings can be compared with GOES >10 MeV threshold crossing to create a contingency table and explore whether the two warning conditions are related. Using these criteria, the raw contingency tables are shown in Table 6.76 and the resulting metrics in Table 6.77. On an event-by-event basis, HESPERIA REleASE ACE 60-Min achieved hits for 25 SEP events, missing only one, while SOHO 60-Min achieved hits for 19 events and missed three. The raw scores report very high Hit Rates of 94% and 88% with extremely low False Alarm Rates of 0.4% and 0.1%, respectively. The raw False Alarm Ratios are still somewhat high at 73% and 42%, but the extreme imbalance of the dataset and very low False Alarm Rates result in very high HSS of 0.42 and 0.70 and TSS of 0.93 and 0.88, respectively. The very small False Alarm Rates also imply that HESPERIA REleASE is an effective predictor for clear conditions – if a forecast has a status of All Clear, it is reliable to trust that the conditions will remain clear at least for the next 20 minutes. Some models have investigated coupling with REleASE to reduce false alarms and this demonstrates that this approach would be an effective strategy.

Figure 6.42 shows HESPERIA REleASE forecasts (black circles) issued every 5

[16] https://hesperia.astro.noa.gr/hesperia-release-alert/

Mixed energies

ACE 60-Min

	Observed		
	Yes	No	Sum
Pred. Yes	296	787	1,083
Pred. No	20	180,231	180,251
Sum	316	181,018	181,334

SOHO 60-Min

	Observed		
	Yes	No	Sum
Pred. Yes	620	441	1,061
Pred. No	87	373,126	373,213
Sum	707	373,567	374,274

Table 6.76: Contingency table comparing HESPERIA REleASE 15.8–39.8 MeV differential protons exceeding 0.1 [MeV cm^2 s sr]$^{-1}$ to GOES >10 MeV integral protons exceeding 10 pfu.

HESPERIA REleASE All Clear	ACE 60-Min Mixed energies ($N = 181,334$)	SOHO 60-Min Mixed energies ($N = 374,274$)
Percent Correct	0.996	0.998
Hit Rate	0.94	0.88
False Alarm Rate	0.0043	0.0012
False Alarm Ratio	0.73	0.42
Bias	3.43	1.5
Threat Score	0.27	0.54
HSS	0.42	0.70
TSS	0.93	0.88

Table 6.77: HESPERIA REleASE ACE 60-Min and SOHO 60-Min raw All Clear metrics comparing HESPERIA REleASE 15.8–39.8 MeV differential protons exceeding 0.1 [MeV cm^2 s sr]$^{-1}$ to GOES >10 MeV integral protons exceeding 10 pfu.

HESPERIA REleASE Advance Warning Time (AWT)	Scoreboard ACE-60 min ($N = 25$)	SOHO-60 min ($N = 19$)
AWT to Observed SEP Start Time		
Median	3.4 hr	1.0 hr
Worst	−2 min	−1 min
Best	19.8 hr	19.8 hr

Table 6.78: HESPERIA REleASE Advance Warning Time
Advance Warning Time for HESPERIA REleASE on the SEP Scoreboard for 15.8–39.8 MeV, 0.1 [MeV cm^2 s sr]$^{-1}$ compared to >10 MeV, 10 pfu.

minutes using ACE electron flux (gray line) or SOHO electron flux (not plotted) as input. The HESPERIA REleASE 60-Min points are plotted 1 hour into the future and thus were issued 1 hour prior to the observational data appearing at the same time. The GOES measurements are colored lines with >10 MeV in red. The electron flux produces a similar time profile to the proton flux and the HESPERIA REleASE forecasts closely follow the behavior of the GOES integral fluxes. Figures 6.43 and 6.44 plot the predicted proton flux compared to the maximum observed >10 MeV proton flux in the forecast prediction window. It is evident that the forecasts and measurements are highly correlated. A bias is present, which is expected due to the differences in their energy ranges. The ACE-derived predictions have a tendency towards false alarms, as evidenced by the numerous high predictions that exceed the 0.1 flux threshold while GOES flux remains low. SOHO-derived predictions are much cleaner, with only a few false alarms exceeding the 0.1 threshold during quiet conditions. Recent efforts from the REleASE team have incorporated Type III radio bursts into the prediction scheme to further reduce false alarms and this has shown to be a promising approach.

HESPERIA REleASE is designed to predict the onset phase of SEP events. The maximum predicted value during the onset phase can be interpreted as a prediction for peak flux. Figures 6.45 and 6.46 plot a comparison of the GOES >10 MeV onset peak to the maximum flux predicted by HESPERIA ReleASE. As seen earlier, there is a bias between the forecasts and predictions, but the values are strongly correlated.

These results indicate that, although it does not predict the absolute value of the GOES proton flux, HESPERIA REleASE can inform how the GOES proton flux time profile is likely evolve in the next few hours.

Advance warning time was calculated by identifying the first HESPERIA REleASE forecast above 0.1 [MeV cm^2 s sr]$^{-1}$ ahead of an SEP event and comparing the forecast issue time to the GOES observed >10 MeV, 10 pfu threshold crossing time. The HESPERIA REleASE advance warning time analysis produce the results in Table 6.78. The full spread of AWT values are plotted in Figure 5.45. HESPERIA REleASE consistently provides advance warning of an hour or more prior to GOES >10 MeV threshold crossings.

In summary,

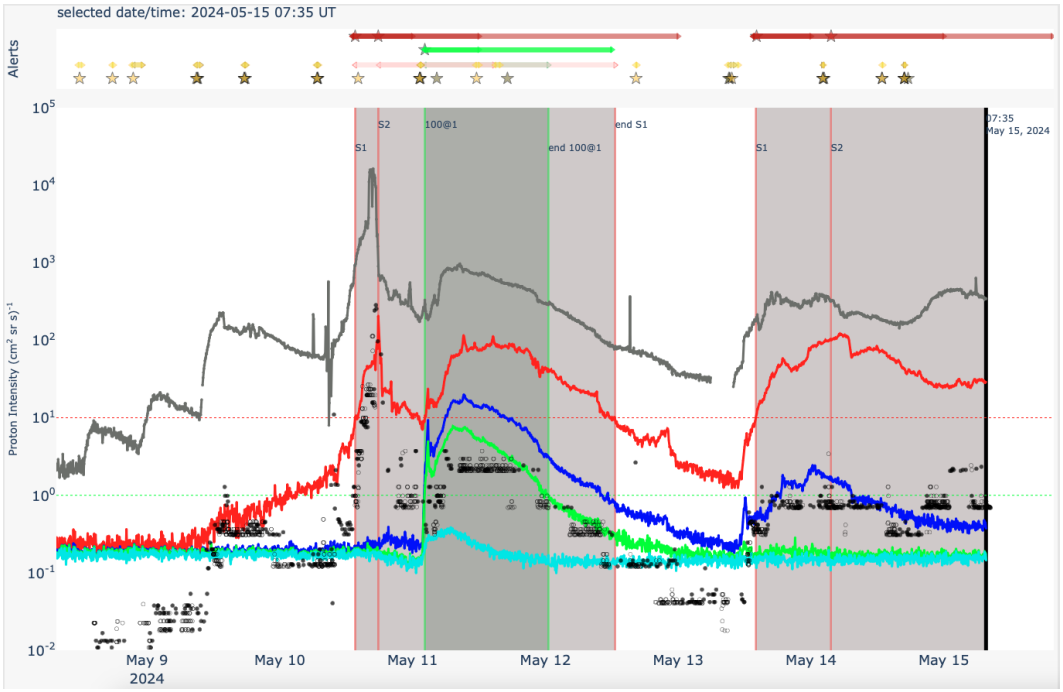


Figure 6.42: HESPERIA REleASE forecasts (black circles) on the SEP Scoreboard during the May 2024 Mother's Day/Gannon Storms, plotted alongside GOES integral channels (colors) and ACE electrons(grey). The HESPERIA REleASE 60-Min points are plotted 1 hour into the future and thus were issued 1 hour prior to the observational data appearing at the same time.

- A comparison between HESPERIA REleASE's warning conditions for 15.8–39.8 MeV and GOES >10 MeV threshold crossings show very high skill in real time on the SEP Scoreboards. HESPERIA ReleASE achieved very high Hit Rates of 94% and 88% with extremely low False Alarm Rates of 0.4% and 0.1%, when using ACE or SOHO as input, respectively. False Alarm Ratios are 73% and 42%. This results in HSS = 0.42 for ACE and 0.70 for SOHO-driven forecasts.

- HESPERIA REleASE forecasts using SOHO/EPHIN electron fluxes as input are rather clean. ACE/EPAM, however, suffers from recurrent high electron-intensity outliers (see May 10, May 12, May 15 in Figure 6.42, also high-prediction 'clouds' in Figures 6.43 and 6.45) and is thus more likely to produce false alarms. Due to the input data quality, SOHO-based forecasts result in higher skill forecasts compared to ACE/EPAM-based forecasts.

- When comparing the time that HESPERIA REleASE issues a warning for its warning condition with the observed >10 MeV threshold crossing time, the higher skill HESPERIA REleASE SOHO-60 min forecast has a median AWT of 1 hour while ACE-60 min achieves a median AWT of 3.5 hours.

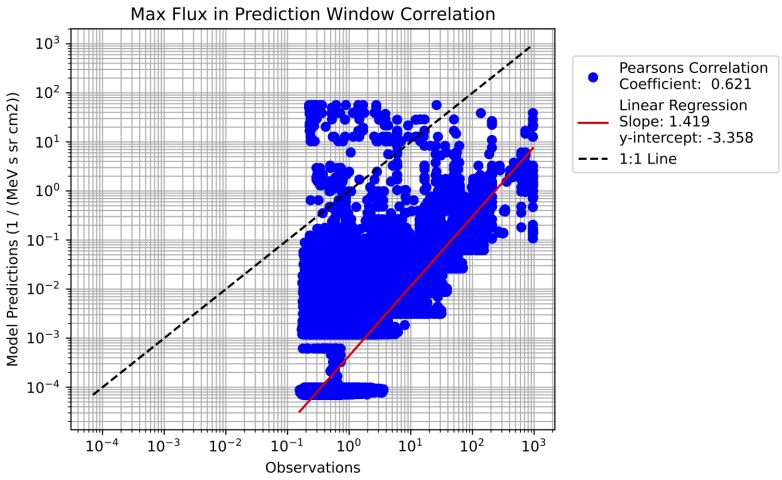


Figure 6.43: Maximum flux in the prediction window for REleASE ACE-60 Min 15.8–39.8 MeV compared to GOES >10 MeV fluxes for all forecasts on the SEP Scoreboards. It is not expected that the correlation would fall on the 1:1 line because the forecast and observations are for different proton energies.

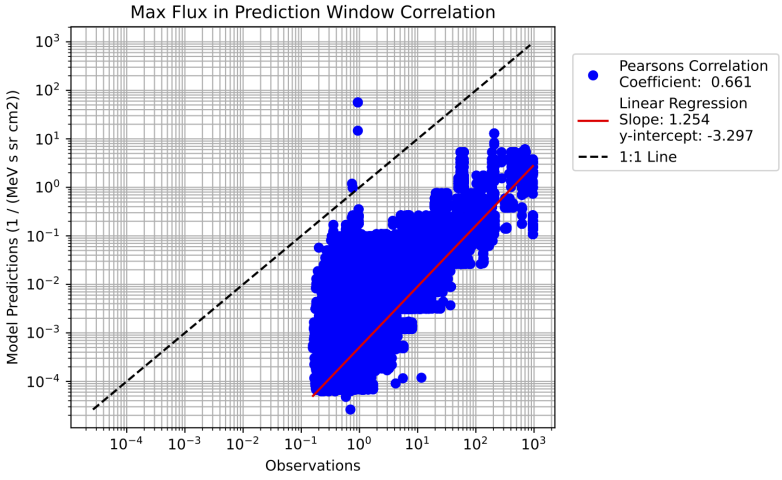


Figure 6.44: Maximum flux in the prediction window for REleASE SOHO-60 Min 15.8–39.8 MeV compared to GOES >10 MeV fluxes for all forecasts on the SEP Scoreboards. It is not expected that the correlation would fall on the 1:1 line because the forecast and observations are for different proton energies.

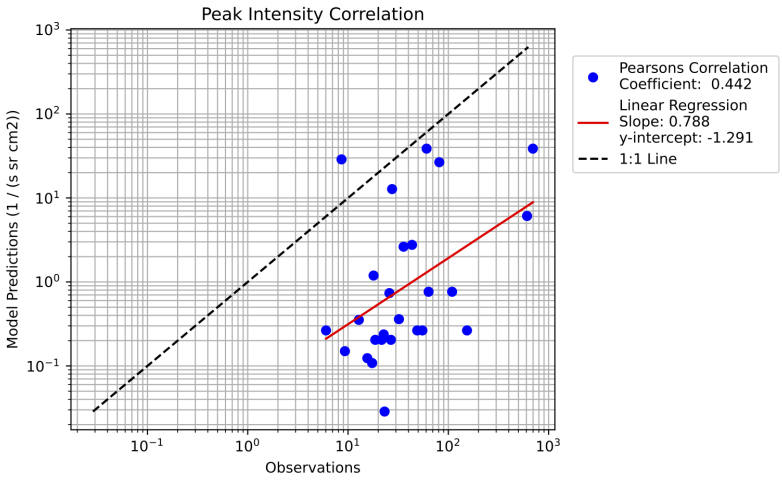


Figure 6.45: Maximum 15.8–39.8 MeV proton flux predicted by REleASE ACE-60 Min compared to GOES >10 MeV onset peak fluxes for SEP events on the SEP Scoreboards. It is not expected that the correlation would fall on the 1:1 line because the forecast and observations are for different proton energies.

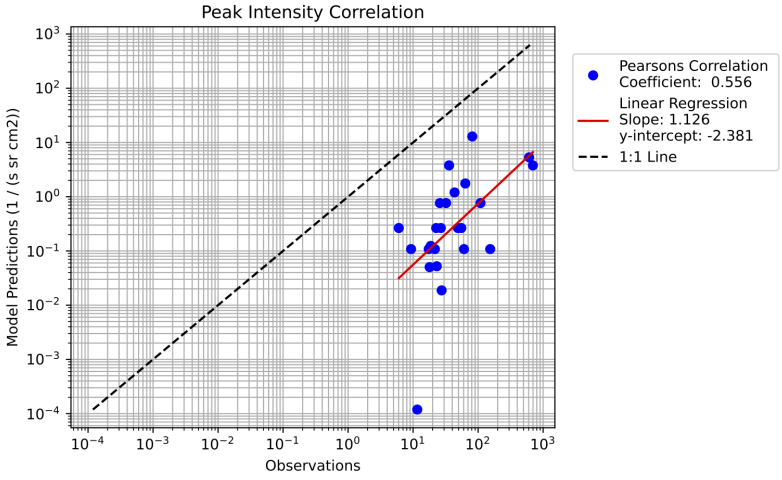


Figure 6.46: Maximum 5.8–39.8 MeV proton flux predicted by REleASE SOHO-60 Min compared to GOES >10 MeV onset peak fluxes for SEP events on the SEP Scoreboards. It is not expected that the correlation would fall on the 1:1 line because the forecast and observations are for different proton energies.

7 Cross-Model Comparisons

7.1 24-hour SPE Forecasting Models

In this section we compare models issuing a 24-hour forecast for the occurrence of an SPE. We include MAG4 LOS_r and SHARP_HMI (Section 6.2) and MagPy SHARP_HMI (Section 6.3). The additional MAG4 variants were omitted because we lack SEPVAL results for them, and the GSU model was omitted because its performance was not competitive (see Section 6.4). Each of these automated models is compared to the human-driven SWPC (Section 6.1), our base reference point of comparison.

	SEPVAL			
Model	MAG4	MAG4	MagPy	SWPC
Variant	LOS_r	SHHMI	SHHMI	Day-1
N	897	547	2003	58
Brier Score	0.19	**0.22**	0.36	0.41
Brier Skill Score	0.00	−0.03	0.04	**0.12**
Brier (SEP)	0.92	0.98	0.90	**0.80**
Median P (SEP)	**0.03**	0.01	**0.03**	**0.03**
Area Under the Curve	**0.67**	0.56	0.56	0.48

	Scoreboard			
Model	MAG4	MAG4	MagPy	SWPC
Variant	LOS_r	SHHMI	SHHMI	Day-1
N	27150	19288	11117	2135
Brier Score	0.03	0.03	0.05	**0.02**
Brier Skill Score	−0.037	−0.042	**−0.01**	−0.25
Brier (SEP)	0.92	0.96	0.91	**0.69**
Median P (SEP)	0.03	0.01	0.03	**0.13**
Area Under the Curve	0.65	0.58	0.56	**0.78**

Table 7.1: 24-hour SPE probability metrics from SEPVAL and Scoreboard forecasts. SHHMI refers to the SHARP_HMI variant. Best values for the metric are in boldface.

Table 7.1 shows the probability metrics for the selected models and SWPC. SWPC has the overall best probability performance, best captured by its high AUC and high median probability for forecasts matching SPEs. It has the worse Brier Skill Scores, likely due to their high probability estimates (see Figure 6.2). The remaining models have very similar performance, with MAG4 LOS_r showing the next highest performance, followed by MagPy. MagPy and MAG4 SHARP_HMI are the most directly comparable in terms of their design. Comparing those, we see that MagPy has a slightly higher AUC in the SEPVAL dataset, but slightly lower on the Scoreboard. The Brier scores for MagPy are worse, likely due to its apparently higher (but more reliable) probability scale, see Figures 6.5 vs. 6.10.

Table 7.2 shows All Clear metrics for the selected model set. MAG4 LOS_r,

SEPVAL				
Model	MAG4	MAG4	MagPy	SWPC
Variant	LOS_r	SHHMI	SHHMI	Day-1
P Threshold	1%	1%	7%	10%
Imbalance	0.70	0.73	1	1
N	56	57	62	59
Percent Correct	0.55	**0.56**	0.55	0.52
Hit Rate	**0.67**	0.48	0.67	0.27
False Alarm Rate	0.61	0.33	0.59	**0.21**
False Alarm Ratio	0.39	0.33	0.44	**0.43**
Bias	**1.09**	0.73	1.18	0.47
Threat Score	0.47	0.39	**0.44**	0.22
HSS	0.06	**0.14**	0.08	0.06
TSS	0.06	**0.15**	0.08	0.06

Scoreboard				
Model	MAG4	MAG4	MagPy	SWPC
Variant	LOS_r	SHHMI	SHHMI	Day-1
P Threshold	1%	1%	19%	10%
Imbalance	35.6	36.2	21.5	62.9
N	1275	1233	506	2135
Percent Correct	0.42	0.83	**0.91**	**0.91**
Hit Rate	**0.80**	0.38	0.09	0.50
False Alarm Rate	0.59	0.15	**0.05**	0.09
False Alarm Ratio	0.96	0.93	**0.91**	**0.91**
Bias	21.8	5.8	**1.22**	5.82
Threat Score	0.04	0.06	0.04	**0.08**
HSS	0.02	0.07	0.04	**0.12**
TSS	0.21	0.23	0.03	**0.41**

Table 7.2: 24-hour SPE All Clear metrics from SEPVAL and Scoreboard forecasts. SHARP refers to the SHARP_HMI variant. Best values for the metric are in boldface.

with a low 1% probability threshold, has the highest Hit Rate of the models, hitting 67% and 80% of the events in SEPVAL and the SEP Scoreboards, respectively. This comes at the cost of a high False Alarm Rate, indicating a higher propensity towards false alarms compared to the other models. MagPy performed poorly on the Scoreboard, hitting only 9% of events, but performed comparably to all other models in the SEPVAL set when a lower 7% probability threshold was applied (compared to 19% on the Scoreboard). This strongly indicates that a reduced threshold should be deployed into production; indeed the threshold was adjusted to 15% in December of 2024 on the SEP Scoreboards. SWPC All Clear performance at the 10% threshold achieves a Hit Rate of 50% and the highest TSS compared to the other models, but our study in choosing the threshold showed that a lower False Alarm Ratio is

not achievable. It should be noted that SWPC results are over a longer and more imbalanced dataset, making low False Alarm Rate and high HSS more difficult to achieve.

MagPy and SWPC are directly comparable for SEPVAL, however, and it is seen that MagPy had similar probability performance and in All Clear slightly outperformed the forecasting office in all but False Alarm Rate, but achieving significantly higher Hit Rate (67% compared to 27%). This is an interesting result, and invites speculation as to whether the purely analytic approach of MagPy offers some advantages over human decision making for this specialized subset of relatively active conditions for both non-event and event periods. Another possibility is that the validation approach of collapsing a series of ∼24 hourly forecasts before the event/non-event flare or CME into a single result was to the model's benefit, considering that SWPC only issues one forecast per day. A deeper analysis is needed to confirm that the comparison is truly fair and that the result is not a statistical fluke. We do note, however, that it is a practical advantage in general of the automated models that they can offer updated forecasts at a higher cadence.

In summary,

- SWPC Day-1 has the highest performance for 24-hour SPE forecasts. Models which consider only magnetic field properties as inputs may be improved by considering other indicators that the forecasting office uses.
- MagPy SHARP_HMI has comparable to slightly superior performance to its predecessor, MAG4 SHARP_HMI. However, MagPy SHARP_HMI performance is below that of MAG4 LOS_r, indicating that the lower fidelity LOS observations and LOS analysis techniques produce superior results.
- Comparing MagPy to SWPC with the SEPVAL challenge set shows that MagPy had similar probability performance and in All Clear, slightly outperforming the forecasting office in all but False Alarm Rate, but achieving significantly higher Hit Rate (67% compared to 27%).

7.2 SEPSTER and SEPSTER2D

One of the goals of the SEPVAL challenge was to run models using similar workflows and the same inputs to allow for model cross-comparisons. SEPSTER and SEPSTER2D ingest the exact same inputs and share a common fundamental assumption that SEP peak flux can be predicted as a function of CME source longitude and CME speed, making it particularly appropriate to compare these models side-by-side. But, as a caveat, the N forecasts for the two models are not exactly the same (see Table 7.3): both models provided 60 forecasts for the >100 MeV channel, but SEPSTER gave >10 MeV forecasts for 61 event periods while SEPSTER2D provided 59. Even though both models provided 60 forecasts for the >100 MeV channel out of the total 63 challenge periods, the 3 missing forecasts in each may not be for the same event periods. Despite this caveat, the metrics presented below are robust enough to draw conclusions.

Table 7.4 shows the SEPVAL metrics for the two models. Despite having similar percent correct for each energy (>10 MeV: 74% and 71%, >100 MeV: 82% and 78%), the other metrics show quite different performance. SEPSTER2D has higher Hit Rates than SEPSTER, but also higher False Alarm Rates and Ratios, meaning that SEPSTER2D tends to overpredict the occurrence of SEP events. This overprediction leads to SEPSTER2D achieving the same or higher threat scores, but lower HSS and TSS.

SEPSTER (Parker Spiral)

> 10 MeV

	Observed		
	Yes	No	Sum
Pred. Yes	20	5	25
Pred. No	11	25	36
Sum	31	30	61

> 100 MeV

	Observed		
	Yes	No	Sum
Pred. Yes	7	2	9
Pred. No	9	44	53
Sum	16	44	60

SEPSTER2D

> 10 MeV

	Observed		
	Yes	No	Sum
Pred. Yes	27	14	41
Pred. No	3	15	18
Sum	30	29	59

> 100 MeV

	Observed		
	Yes	No	Sum
Pred. Yes	8	5	13
Pred. No	8	39	47
Sum	16	44	60

Table 7.3: SEPSTER (Parker Spiral) and SEPSTER2D contingency tables for the SEPVAL dataset.

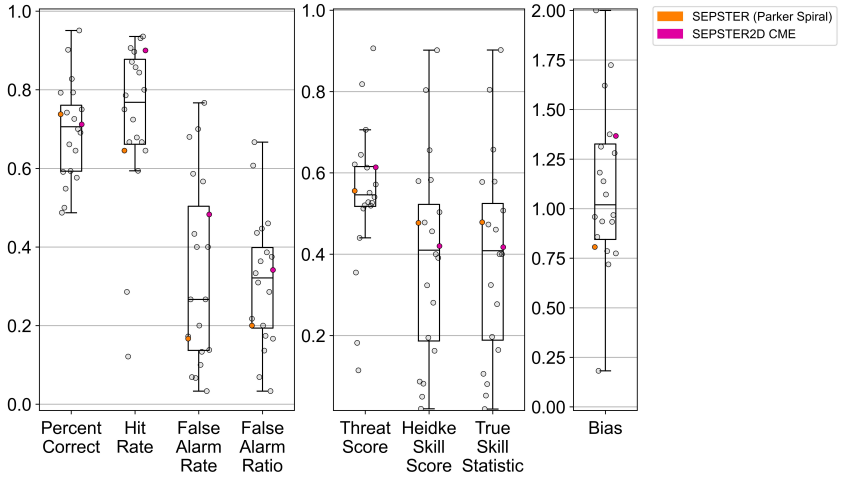
Since these two models share a similar workflow and baseline assumption, we can compare how well this assumption performs in comparison to the other models in the SEPVAL challenge. Figure 7.1 shows the box plot distributions of the various All Clear metrics, with SEPSTER and SEPSTER2D emphasized. SEPSTER is in the better performing quartiles for all metrics except Hit Rate, where it is in the bottom 25% of models. SEPSTER2D also scores in the better performing quartiles for all metrics except for False Alarm Rate and Ratio. Despite a systematic underprediction by SEPSTER and a systematic overprediction by SEPSTER2D, they still achieve better than median skill compared to the group of participating SEPVAL

SEPVAL All Clear	SEPSTER (Parker Spiral)		SEPSTER2D	
	> 10 MeV ($N = 62$)	> 100 MeV ($N = 60$)	> 10 MeV ($N = 59$)	> 100 MeV ($N = 60$)
Percent Correct	0.74	0.82	71	0.78
Hit Rate	0.65	0.44	0.90	0.50
False Alarm Rate	0.17	0.04	0.48	0.11
False Alarm Ratio	0.20	0.22	0.34	0.38
Threat Score	0.56	0.39	0.61	0.38
HSS	0.48	0.46	0.42	0.41
TSS	0.48	0.39	0.42	0.39

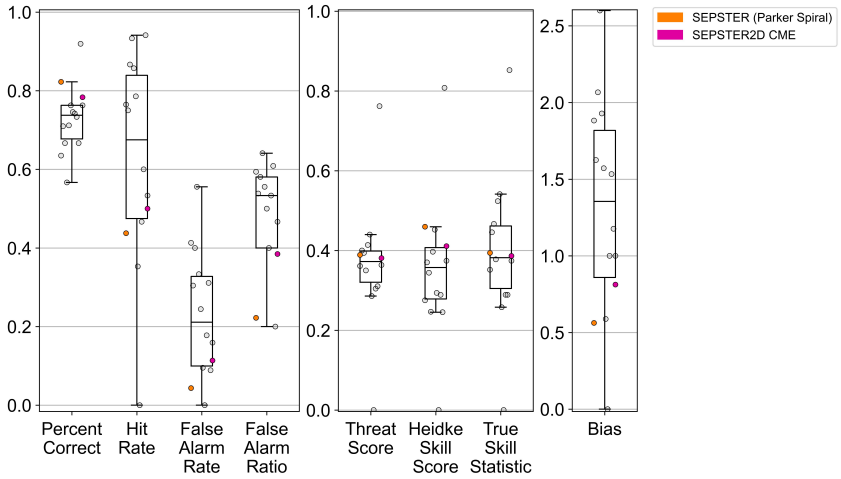
Table 7.4: SEPSTER (Parker Spiral) and SEPSTER2D All Clear Metrics for SEPVAL

models, indicating that the approach has merit.

Comparisons of onset peak flux predictions to observations for the two models can be seen in Figure 7.2. As described in the individual model summaries, SEPSTER (Parker Spiral) was initially trained to predict the onset peak flux whereas SEPSTER2D was trained to predict the maximum peak flux, which could include enhancements from ESPs. This will lead to slightly different predictions when given the same input parameters, with SEPSTER2D expected to give slightly larger peak flux values. For SEPSTER and SEPSTER2D, the onset peak metrics include forecasts for all observed SEP events, whether the two models predicted a peak above threshold or not. This is different from SEPMOD and iPATH, whose metrics include only the subset of observed SEP events that were correctly predicted to cross threshold. For this reason, it is more appropriate to compare the two SEPSTERs to each other. The >10 MeV correlation plot 7.2a, shows both models have a positive correlation with observations, with SEPSTER2D having higher predicted values overall as expected. For the >100 MeV energy channel, the models exhibit different performance, with SEPSTER (Parker Spiral) having a flatter best fit line (but still positively sloped), showing less correlation between its predictions and observations.

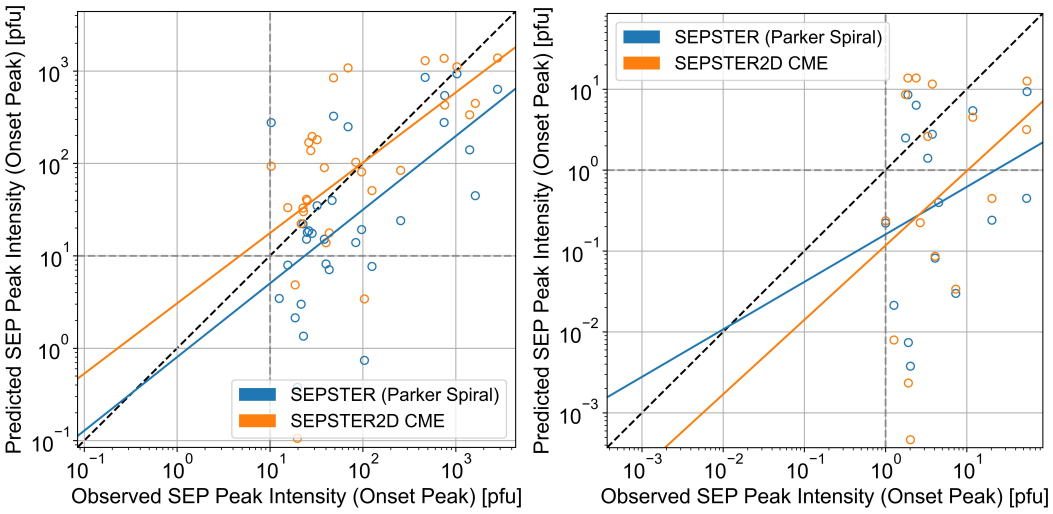


(a) > 10 MeV



(b) >100 MeV

Figure 7.1: Box plot distributions of All Clear metrics for all models participating in the SEPVAL challenge. All models but SEPSTER and SEPSTER2D are de-emphasized. Top is >10 MeV and bottom is >100 MeV.



(a) > 10 MeV (b) >100 MeV

Figure 7.2: Scatter plots of Observed Onset Peak flux against Predicted Onset Peak flux for SEPSTER (Parker Spiral) and SEPSTER2D for the SEPVAL dataset, with best fit lines demonstrating correlation between observations and predictions. Left is >10 MeV and Right is >100 MeV.

7.3 SEPMOD and iPATH

Both iPATH and SEPMOD are physics-based SEP models that aim to model the transport of particles in the inner heliosphere produced by eruptions from the Sun. They share the same baseline assumptions that particles are accelerated via diffusive shock acceleration, where particles are energized by multiple repeated crossings of the shock front before eventually escaping the shock. However, the details of these models and how they simulate these phenomena are very different. Both of these models were run in their real-time configurations for SEPVAL using the same input parameters and without fine-tuning the results. For this reason, it is reasonable to make direct comparisons between the two. It should be noted that SEPMOD provided 63 forecasts for SEPVAL while iPATH provided 60, so their datasets are slightly different, but we can still draw overall conclusions.

iPATH

> 10 MeV

	Observed		
	Yes	No	Sum
Pred. Yes	20	4	24
Pred. No	11	25	36
Sum	31	29	60

> 100 MeV

	Observed		
	Yes	No	Sum
Pred. Yes	8	7	15
Pred. No	7	37	44
Sum	15	44	59

SEPMOD

> 10 MeV

	Observed		
	Yes	No	Sum
Pred. Yes	19	4	23
Pred. No	13	26	39
Sum	32	30	62

> 100 MeV

	Observed		
	Yes	No	Sum
Pred. Yes	6	4	10
Pred. No	11	38	49
Sum	17	42	59

Table 7.5: iPATH and SEPMOD contingency tables for the SEPVAL dataset.

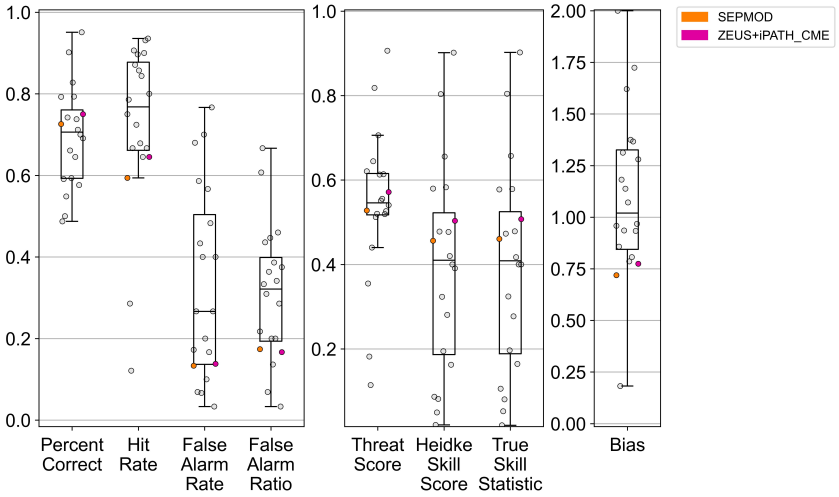
Table 7.5 shows the contingency tables for both models. The All Clear metrics for the two models are shown in Table 7.6. Similar performance (within a few percent) is shown in the models for percent correct, False Alarm Rate, and False Alarm Ratio with respect to each energy, noting that False Alarm Ratio is higher for both in the >100 MeV channel. Due to SEPMOD's lower Hit Rate, iPATH achieves higher values in the skill scores for both energies. Figure 7.3 shows box plot distributions with SEPMOD and iPATH (highlighted) compared to all SEPVAL models. These distributions show that, despite the differences in the physics implementations, their performance across most metrics is similar (same quartile) and near or above the median of the models for >10 MeV. A difference is seen in Figure 7.3b for >100 MeV performance where SEPMOD has lower skill than nearly all of the other models, whereas iPATH remains near the center of the pack and, notably, for bias has perfect score at 1.0, meaning that it does not tend more towards false alarms or misses. This means that for the higher energies, iPATH is better at predicting when there will be an ESPE event than SEPMOD. Otherwise, it is promising that these physics-based models perform in the highest quartile for the >10 MeV HSS and TSS metrics.

As both of these models predict time profiles, we can look at how the metrics compare for the peak flux metrics. We will focus on onset peak flux, since the onset

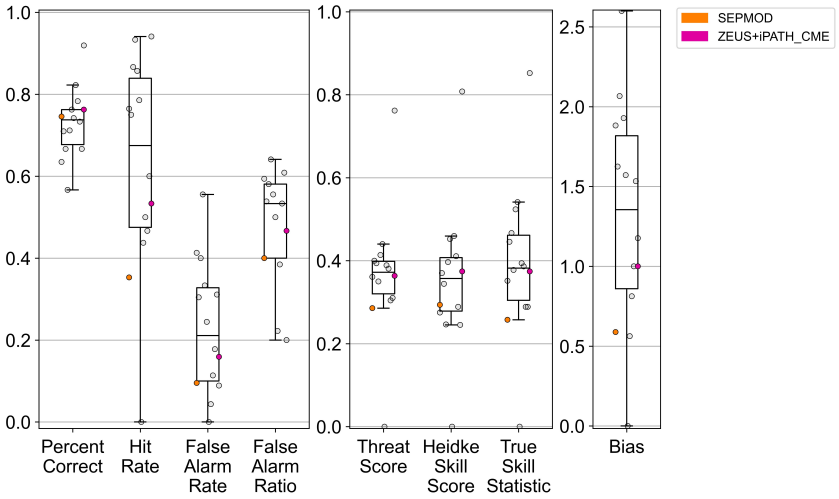
SEPVAL	iPATH		SEPMOD	
All Clear	> 10 MeV	> 100 MeV	> 10 MeV	> 100 MeV
	($N = 60$)	($N = 59$)	($N = 62$)	($N = 59$)
Percent Correct	0.75	0.76	0.73	0.75
Hit Rate	0.65	0.53	0.59	0.35
False Alarm Rate	0.14	0.16	0.13	0.10
False Alarm Ratio	0.17	0.47	0.17	0.40
Threat Score	0.57	0.36	0.53	0.29
HSS	0.50	0.37	0.46	0.29
TSS	0.50	0.37	0.46	0.26

Table 7.6: iPATH and SEPMOD All Clear Metrics for SEPVAL

peak is calculated from both models using FetchSEP following the same approach applied to the observations. In this case, the onset peak metrics will include only the subset of observed SEP events associated with forecasts that correctly predicted the flux would cross threshold (i.e., only hits). Figure 7.4 shows scatter plots of the observed onset peak flux and the predicted onset peak flux for the two models for the SEPVAL dataset. For the >10 MeV channel, both models show positive slope, whereas for the >100 MeV, both SEPMOD and iPATH have a nearly flat or negative slope. For >100 MeV, iPATH over-predicts for small ESPE events and near the 1:1 line for larger events, whereas SEPMOD has a lot of scatter with no particular trending. We interpret this to mean that both models do not show any skill in determining the event-to-event variability of ESPE events. For >10 MeV, the models are more successful at capturing trending from event to event, however, as pointed out in earlier sections for all models, the peak flux predictions may be 1 to 2 or more orders of magnitude from the 1:1 line.

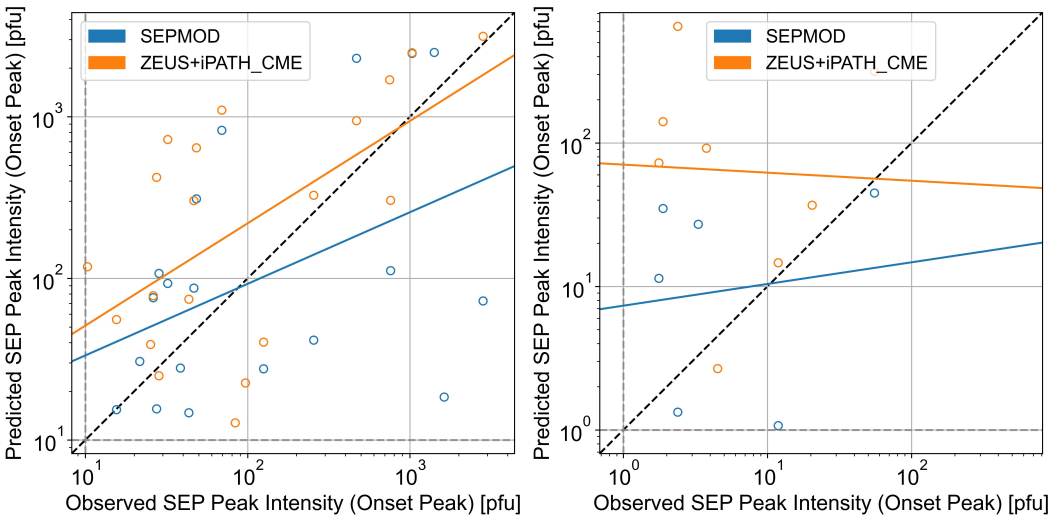


(a) > 10 MeV



(b) > 100 MeV

Figure 7.3: Box plot distributions of All Clear metrics for all models participating in the SEPVAL challenge. All models but SEPMOD and iPATH are de-emphasized. Top is >10 MeV and bottom is >100 MeV.



(a) >10 MeV (b) >100 MeV

Figure 7.4: Scatter plots of Observed Onset Peak flux against Predicted Onset Peak flux for SEPMOD and iPATH for the SEPVAL dataset, with best fit lines demonstrating correlation between observations and predictions. Left is >10 MeV and Right is >100 MeV.

7.4 SWPC Warnings and Post-eruptive Forecasting Models

SWPC Warnings are issued following conditions that may cause the >10 MeV protons to rise above threshold, such as flares, CMEs, or observed increases in particle flux at Earth, so these forecasts are considered to fall in the post-eruptive category. This section compares the performance of SWPC Warnings to the post-eruptive models active on the SEP Scoreboards: SAWS-ASPECS Flare, ENLIL+SEPMOD, SEPSTER (Parker Spiral), SEPSTER2D, SPRINTS, UMASEP, and ZEUS+iPATH. This comparison will use the All Clear metrics from both the SEPVAL challenge and the SEP Scoreboards for the >10 MeV protons.

SWPC Warnings are only issued when a SWPC forecaster believes that the proton environment will cross operational thresholds in the very near future. There is no analog "all clear" forecast, rather the lack of a Warning indicates the forecaster believes protons will remain below threshold. For SEPVAL, as there were discrete challenge periods specified, the lack of Warning for non-event periods was taken as a correct negative while the lack of a Warning prior to the start of a SEP event was counted as a miss. For the SEP Scoreboards, we did not attempt to count the lack of Warnings as correct negatives, thus restricting the calculation of All Clear metrics to those that include only hits, misses, and false alarms. Percent correct for SWPC Warnings on the SEP Scoreboards is $Hits/(Hits + Misses + False\ Alarms)$.

For SEPVAL, SWPC issued one false alarm and five misses, resulting in a Hit Rate of 0.85 and FAR of 0.034, with other metrics shown in Table 7.7. The only model that achieved a similarly good performance was UMASEP-10, with a Hit Rate of 0.94 and FAR of 0.033. SEPSTER2D also has a high Hit Rate of 0.90, however its FAR of 0.341 indicates that it issues many more false alarms compared to SWPC. Other models achieve Hit Rates in the ∼0.60 to 0.75 range, but all have much larger FAR than SWPC.

For the SEP Scoreboards, it should be noted that each model (including SWPC) has been running for varying time frames with different numbers of SEP events and somewhat different, but realistic, climatological imbalances in their datasets. SWPC Warning had 0.71 percent correct with a False Alarm Ratio of 0.29 (12 false alarms and 29 hits from a total of 41 forecast periods, deoverlapped). Notably, SWPC's Hit Rate is 1.0, meaning that they issued forecasts ahead of every SEP event that occurred during the Scoreboard time period. No other model, as shown in Table 7.8, has such good statistics for this real-time dataset. In real time for SC 25, the models are not able to achieve as high a Hit Rate and as low a False Alarm Ratio as SWPC forecasters.

These All Clear results imply that the information used by SWPC forecasters coupled with human intuition outperforms models in a real-time environment like Solar Cycle 25. For more intense time periods dominated by larger flares and faster CMEs, like those selected for SEPVAL, only UMASEP-10 matches or exceeds the skill of SWPC's Warning product. Overall, post-eruptive model performance on the SEP Scoreboards for All Clear cannot match SWPC's Warnings, however many of these models provide other types of predictions (peak flux, time profile) which are useful in operations for situational awareness.

SEPVAL	Percent Correct	Hit Rate	False Alarm Rate	False Alarm Ratio	TSS	HSS
SWPC Warning	0.90	0.85	0.03	0.03	0.82	0.81
ASPECS flare	0.66	0.72	0.40	0.36	0.32	0.32
ENLIL+SEPMOD	0.73	0.59	0.13	0.17	0.46	0.46
SEPSTER (PS)	0.74	0.65	0.17	0.20	0.48	0.48
SEPSTER2D	0.71	0.90	0.48	0.34	0.42	0.42
SPRINTS	0.70	0.67	0.27	0.29	0.40	0.40
UMASEP-10	0.95	0.94	0.03	0.03	0.90	0.90
ZEUS+iPATH	0.75	0.65	0.14	0.17	0.51	0.50

Table 7.7: All Clear metrics for post-eruptive models in SEPVAL.

SEP Scoreboards	Percent Correct	Hit Rate	False Alarm Rate	False Alarm Ratio	TSS	HSS
SWPC Warning	0.71	1.0	-	0.29	-	-
ENLIL+SEPMOD	0.94	0.26	0.048	0.90	0.21	0.12
SEPSTER (PS)	0.98	0.62	0.018	0.64	0.61	0.44
SEPSTER2D	0.69	0.94	0.33	0.84	0.61	0.19
SPRINTS	0.99	0.13	0.002	0.74	0.13	0.14
UMASEP-10	0.96	0.69	0.031	0.63	0.65	0.46
ZEUS+iPATH	0.92	0.60	0.063	0.69	0.53	0.37

Table 7.8: All Clear metrics for post-eruptive models on the Scoreboard.

SWPC is the primary provider of space weather forecasts for SRAG, therefore it is interesting to quantify the advance warning provided by SWPC's forecasts in comparison with models, listed in Table 7.9. AWT described the amount of time prior to an observed proton flux crossing operational thresholds that a forecast is issued. A positive AWT indicates that the forecast was issued before the threshold crossing, while a negative AWT indicates the forecast was issued afterwards. For the SEP Scoreboard periods covering 29 SEP events, SWPC Warnings were issued 1.03 hours (median) before observed thresholds were crossed. SWPC Warning forecasts thus couple high skill with an early warning of an hour or more AWT for at least half of the SEP events in SC 25, providing a clear benefit to operations. UMASEP-10 is the only model that provides similar AWT with high skill. The other models do not provide any AWT to SEP start time, and these all use M2M DONKI CME parameters, which involve hours of waiting to downlink LASCO imagery and a human in the loop to analyze any observed CME. It should be noted that HESPERIA-REleASE SOHO-60 min (see Section 6.12) demonstrated similar high skill and AWT, however the forecasts are made for a different set of warning conditions, and while they appear to be analogous, are not directly comparable to the energy channels and thresholds forecast by SWPC.

Model	N_{SEP}	AWT to Start Time (hours)
SWPC Warning	29	1.03
ENLIL+SEPMOD	8	−3.45
SEPSTER (PS)	15	−0.61
SEPSTER2D	27	−0.55
SPRINTS	5	1.82
UMASEP-10	23	0.74
ZEUS+iPATH	11	−2.93

Table 7.9: Median Advance Warning Time for the >10 MeV, 10 pfu threshold crossing for post-eruptive models on the SEP Scoreboard. Positive times indicate that forecasts were issued before the observed proton fluxes crossed threshold.

8 Conclusions & Recommendations

8.1 State of the Art in SEP Forecasting Models

With the proliferation of SEP forecast models, the development of real-time forecast tools through NASA's ISEP project, and the need for a fast response to changing space weather conditions during crewed Artemis missions, it has become imperative to quantify individual SEP model performance and define the state-of-the-art performance of SEP model forecasting.

The SPHINX Validation Framework is an automated, generalized validation program developed to evaluate forecasts from SEP prediction models. It was created to evaluate SEP model performance in a systematic, consistent, and fully reproducible manner. The following results were calculated using this framework as applied to NOAA SWPC forecasting, the SEPVAL community challenge, and the real-time SEP Scoreboards.

The state of the art is defined here as the median scores derived from the SEPVAL and the SEP Scoreboards model groups. When there are enough participating models in a category, a "target quartile" is provided to indicate scores that achieve better performance than 75% of models participating in these efforts. The target quartile is chosen as 75% for metrics that should be maximized, like Hit Rate, or 25% for metrics that aim to be minimized, like False Alarm Rate.

The metrics are organized by "pre-eruptive" and "post-eruptive" models:

- Pre-eruptive models answer the question: "Will an eruption and SEP occur in the next X hours based on current conditions?" These models use the current conditions in active regions in the photosphere and/or corona to predict whether an eruption and associated SEP event will occur. If an eruption has recently been observed on the Sun, then forecasts from these types of models are interpreted as a prediction for the NEXT possible eruption.

- Post-eruptive models produce a forecast following a solar eruption and answer the question: "Will an SEP occur after THIS eruption?" These models use observational inputs that are produced by the eruption itself, like flares, CMEs, or enhancements in the *in situ* proton and electron flux.

8.1.1 SWPC as a Baseline

NOAA SWPC is the official space weather forecasting service for the United States. SWPC Day-1 forecasts are daily probabilistic proton event forecasts for SPEs, or "S1 Events" using the SWPC solar radiation storm scale. SWPC Day-1 forecasts are issued every day at 22:00 UT and are categorized as pre-eruptive forecasts in this analysis. SWPC Warnings are issued when an SEP event appears to be imminent and are categorized as post-eruptive forecasts. The metrics compiled here give an overview of SWPC's performance for SEPVAL and the SEP Scoreboards as derived by SPHINX, reported in Tables 8.1 and 8.2.

SWPC Day 1 as a Baseline

Metrics	SEPVAL >10 MeV	Scoreboard >10 MeV
All Clear		
Percent Correct	0.52	0.91
Hit Rate	0.27	0.50
False Alarm Rate	0.21	0.09
False Alarm Ratio	0.43	0.91
Bias	0.47	5.82
Threat Score	0.22	0.08
HSS	0.06	0.12
TSS	0.06	0.41
Probability		
Brier Score	0.41	0.02
Brier Skill Score	0.12	−0.25
Area Under the Curve	0.48	0.78
Advanced Warning Time (hours)		
Median AWT to SEP Start Time	-	14.7

Table 8.1: SWPC Day 1 as a baseline.
SWPC Day 1 metrics as a baseline for SEP forecasting. A threshold of 10% (selected to maximize skill) was applied to Day 1 probabilistic forecasts to convert to binary All Clear. Positive AWT values indicate forecasts were issued before the SEP start time.

SWPC Warning as a Baseline

Metrics	SEPVAL >10 MeV	Scoreboard >10 MeV
All Clear		
Percent Correct	0.90	0.71
Hit Rate	0.85	-
False Alarm Rate	0.03	-
False Alarm Ratio	0.03	0.29
Bias	0.88	-
Threat Score	0.82	-
HSS	0.81	-
TSS	0.82	-
Advanced Warning Time (hours)		
Median AWT to SEP Start Time	-	0.93
Shortest AWT to SEP Start Time	-	0.033
Longest AWT to SEP Start Time	-	12.7

Table 8.2: SWPC Warning as a baseline.
SWPC Warning metrics as a baseline for SEP forecasting. The choice not to issue a Warning was interpreted as a "clear" forecast for SEPVAL. Positive AWT values indicate forecasts were issued before the SEP start time.

8.1.2 SEPVAL Challenge

SEPVAL was carried out as a community challenge with the participation of SEP model developers who volunteered their time and effort. An approximately equal sample of SEP events and non-event periods were selected by SRAG, focusing on the types of SEP events and parent eruptions (flares and CMEs) that are relevant to SRAG operations. These results can be viewed as an idealized evaluation of model performance using their default workflows. The inputs are of higher quality than real-time observations and ensured to be complete, the provided model forecasts are as complete as possible, and the extreme imbalance of true climatology is not taken into account. The SEPVAL list of challenge periods is available on Zenodo[17], with additional resources provided on the SEPVAL challenge website hosted by CCMC[18].

A definition of state-of-the-art performance derived from SEPVAL for pre- and post-eruptive models are reported in Tables 8.3 and 8.4, respectively.

SEPVAL Pre-eruptive Models

Metrics	>10 MeV Median	>10 MeV Target Quartile	>100 MeV Median	>100 MeV Target Quartile
All Clear				
Percent Correct	0.55	0.56	-	-
Hit Rate	0.58	0.67	-	-
False Alarm Rate	0.46	0.27	-	-
False Alarm Ratio	0.36	0.33	-	-
Bias	0.91	-	-	-
Threat Score	0.42	0.45	-	-
HSS	0.07	0.10	-	-
TSS	0.07	0.10	-	-
Probability				
Brier Score	0.22	-	-	-
Brier Skill Score	−0.003	-	-	-
Area Under the Curve	0.56	-	-	-

Table 8.3: SEPVAL state-of-the-art metrics for pre-eruptive models derived from SEPVAL median performance. Target quartile values indicate performance better than 75% of models. There are only three >10 MeV Probability models, therefore a target quartile cannot be calculated. No pre-eruptive models forecast for >100 MeV or peak flux at any energy.

[17] https://doi.org/10.5281/zenodo.15020584
[18] https://ccmc.gsfc.nasa.gov/community-workshops/ccmc-sepval-2023/

SEPVAL Post-eruptive Models

Metrics	>10 MeV Median	>10 MeV Target Quartile	>100 MeV Median	>100 MeV Target Quartile
		All Clear		
Percent Correct	0.73	0.79	0.74	0.76
Hit Rate	0.82	0.90	0.68	0.84
False Alarm Rate	0.23	0.14	0.21	0.10
False Alarm Ratio	0.25	0.17	0.53	0.40
Bias	1.02	-	1.35	-
Threat Score	0.56	0.63	0.37	0.40
HSS	0.47	0.58	0.36	0.41
TSS	0.47	0.58	0.38	0.46
		Probability		
Brier Score	0.23	0.18	0.17	0.16
Brier Skill Score	-	-	-	-
Area Under the Curve	0.76	0.85	0.78	0.81
		Onset Peak		
Percent within a factor of 10	0.78	0.84	0.60	0.80
Percent within a factor of 2	0.33	0.35	0.33	0.45
Median Log Error	0.06	-	−0.15	-
Median Absolute Log Error	0.94	0.44	0.89	0.61
Spearman Correlation	0.29	0.38	0.22	0.46
		Maximum Flux		
Percent within a factor of 10	0.76	0.88	0.5	0.77
Percent within a factor of 2	0.33	0.41	0.14	0.40
Median Log Error	−0.15	-	−0.50	-
Median Absolute Log Error	0.86	0.57	0.96	0.61
Spearman Correlation	0.39	0.48	0.24	0.44

Table 8.4: SEPVAL state-of-the-art metrics for post-eruptive models derived from SEPVAL median performance. Target quartile values indicate performance better than 75% of models (when possible).

8.1.3 SEP Scoreboards

The SEP Scoreboards are real-time systems used for applied research and environmental awareness. They are being used by SRAG, which is an operational entity, for environmental awareness and to identify promising models, as well as for envisioning how SEP forecasting could be incorporated into operations or used as a support tool for astronauts. Validation of the SEP Scoreboards represents true forecasting performance, including availability and quality of input data in real time, time to make human-in-the-loop analyses, computational run-time, and inherent model approach. The SEP Scoreboards have aggregated real-time forecasts from 10 different SEP forecasting models for over 4 years. These forecasts represent the true model performance in a real-time setting and their utility for operations. The Scoreboards are available publicly at https://ccmc.gsfc.nasa.gov/scoreboards/sep/.

The state-of-the-art performance derived from the SEP Scoreboards for pre- and post-eruptive models are reported in Tables 8.5 to 8.7. The Advance Warning Time reports the median AWT for the model in that category that provides the most advance warning. AWT must be interpreted in terms of model cadence, prediction window, and performance as the inclination towards producing false alarms may inflate AWT.

SEP Scoreboard Pre-eruptive Models

Metrics	>10 MeV Median	>10 MeV Target Quartile	>100 MeV Median	>100 MeV Target Quartile
All Clear				
Percent Correct	0.55	0.74	-	-
Hit Rate	0.71	0.81	-	-
False Alarm Rate	0.46	0.26	-	-
False Alarm Ratio	0.96	0.94	-	-
Bias	16.9	9.67	-	-
Threat Score	0.04	0.05	-	-
HSS	0.03	0.04	-	-
TSS	0.21	0.24	-	-
Probability				
Brier Score	0.03	0.03	-	-
Brier Skill Score	−0.04	−0.03	-	-
Area Under the Curve	0.61	0.64	-	-

Table 8.5: SEP Scoreboard state-of-the-art metrics for pre-eruptive models derived from SEP Scoreboard median performance. Target quartile values indicate performance better than 75% of models (when possible). No pre-eruptive models on the Scoreboards forecast for >100 MeV or peak flux at any energy.

SEP Scoreboard Pre-eruptive Models

Model	N_{SEP}	HSS	AWT to Start	AWT to Peak
Advanced Warning Time (hours) for >10 MeV Events				
MAG4_SHARP_HMI	13	0.07	19.4	-
Advanced Warning Time (hours) for >100 MeV Events				
No models exist				

Table 8.6: SEP Scoreboard state-of-the-art AWT for pre-eruptive models. AWT for the model with the highest HSS is reported. Positive (negative) values indicate forecasts were issued before (after) the SEP start or peak time.

SEP Scoreboard Post-eruptive Models

Metrics	>10 MeV Median	>10 MeV Target Quartile	>100 MeV Median	>100 MeV Target Quartile
		All Clear		
Percent Correct	0.95	0.98	0.99	0.99
Hit Rate	0.61	0.75	0.18	0.55
False Alarm Rate	0.04	0.02	0.006	0.004
False Alarm Ratio	0.67	0.64	0.90	0.87
Bias	1.80	1.41	2.24	1.31
Threat Score	0.26	0.30	0.06	0.12
HSS	0.28	0.41	0.09	0.20
TSS	0.50	0.61	0.14	0.52
		Probability		
Brier Score	0.006	-	0.002	-
Brier Skill Score	-	-	-	-
Area Under the Curve	0.70	-	0.90	-
		Onset Peak		
Percent within a factor of 10	0.78	0.82	0.63	0.86
Percent within a factor of 2	0.29	0.45	0.07	0.21
Median Log Error	0.12	-	−1.12	-
Median Absolute Log Error	0.56	0.40	1.12	0.49
Spearman Correlation	0.23	0.37	0.14	0.26
		Maximum Flux		
Percent within a factor of 10	0.68	0.75	0.41	0.58
Percent within a factor of 2	0.26	0.32	0	0.08
Median Log Error	−0.54	-	−1.25	-
Median Absolute Log Error	0.68	0.59	1.25	0.88
Spearman Correlation	0.21	0.34	0.40	0.51

Table 8.7: SEP Scoreboard state-of-the-art metrics for post-eruptive models derived from SEP Scoreboard median performance. Target quartile values indicate performance better than 75% of models (when possible). SPRINTS is the only post-eruptive model on the Scoreboards that forecasts probability, thus the probability scores report SPRINTS performance.

SEP Scoreboard Post-eruptive Models

Model	Input	N_{SEP}	HSS	AWT to Start	AWT to Peak
Advanced Warning Time (hours) for >10 MeV Events					
SEPSTER (P. Spiral)	CME	16	0.44	−0.61	4.12
HESPERIA REleASE	particle	19	0.70	1.04	17.2
Advanced Warning Time (hours) for >100 MeV Events					
iPath	CME	3	0.20	−9.95	−2.14
UMASEP-100	particle	5	0.46	0.29	4.10

Table 8.8: SEP Scoreboard state-of-the-art AWT for post-eruptive models. AWT for models with the highest HSS are reported for CME and particle inputs. Results for HESPERIA REleASE are reported for the "SOHO 60-min" version comparing REleASE's internal warning condition to the operational >10 MeV threshold crossing. Positive (negative) values indicate forecasts were issued before (after) the SEP start or peak time.

8.2 Recommendations in Support of Operations-to-Research

The validation results in this report provide some broad conclusions about the current state of SEP forecasting model capabilities and where efforts should be focused to improve performance. Here we highlight some of these messages to provide operations-to-research feedback in support of the R2O2R cycle.

8.2.1 Comparison with the State of the Art

For the first time, a benchmark dataset has been created to enable SEP model cross-comparisons. Model developers are encouraged to use the SEPVAL challenge periods as a benchmark set of events for validation in order to compare scores across models and to identify where their model stands with respect to other models in the field. Scores have been calculated here with respect to operational GOES data to reflect the needs of space radiation operations. Also for the first time, real-time forecasts have been aggregated over multiple years through the SEP Scoreboards and evaluated, providing the true performance of SEP models in an operational setting. The stark differences in metrics between these two datasets demonstrate that there is no substitute for evaluating forecasts truly made in real time. As such, model developers are encouraged to participate in the SEP Scoreboards or to develop their own real-time systems to aggregate forecasts and evaluate real-time performance to compare with other models and the state of the art.

New models should aim for performance greater than the median state-of-the-art performance reported here, while exceeding the "target quartile" indicates high skill. **It should be understood that the state of the art does not indicate sufficient skill for use in operations but does represent current model capabilities.** An increase in skill is required to enable operational decision-making based on model forecasts.

8.2.2 Model Improvements

In this analysis, it is clear that previous modeling efforts have focused on the NOAA SWPC definition of an SEP event in the >10 MeV energy channel, as many models show skill in forecasting at these energies. However space radiation operations are primarily concerned with enhancements for >100 MeV particles. Models showed very poor skill in forecasting these energetic events and it is apparent that these rarer events have been an afterthought in the model development process. Significant improvements are needed to forecast energetic >100 MeV SEP events and concentrated efforts towards that goal is identified as a priority for future model development.

A major takeaway of these validation efforts is that the extreme imbalance of SEP events compared to clear periods implies that even very small False Alarm Rates are too high, resulting in more false alarms than hits or even the number of observed SEP events. UMASEP-10 results show that even a 3% False Alarm Rate produced more false alarms (41) than the number of observed SEPs (35) over 4.5 years during the rise of SC 25. UMASEP-10 showed very good performance with 29

hits, but this still resulted in a high False Alarm Ratio of 63%. In general, models must aim for high Hit Rates while reducing false alarms. To achieve high skill in an imbalanced, climatological scenario, False Alarm Rates must be on the order of a few percent while False Alarm Ratios must be less than 50% such that the false alarms do not outnumber hits. Significant improvements are needed, especially for >100 MeV forecasts. Development of strategies to achieve high Hit Rates while reducing false alarms is identified as a priority to increase reliability for operations.

HSS is sensitive to both the Hit Rate and False Alarm Ratio, making it a good choice for optimization of model performance (e.g., for development of machine learning models) compared to other widely used metrics such as True Skill Score (TSS). The numerical value of HSS, however, is extremely sensitive to dataset imbalance and is low when the False Alarm Ratio is high on imbalanced datasets. The use of numerical values for this metric as a requirement must be accompanied by conditions on the dataset for which it is measured. Optimization of HSS for a given dataset, however, should result in skill that is calibrated towards space radiation operations needs.

Validation results for probability forecasts show that there are simple modifications that could improve aspects of model performance. In particular, inspection of ROC curves for models forecasting event probability reveals that there is room for tuning and optimization of Hit Rate/False Alarm Rate by choosing an alternative probability threshold.

Pre-eruptive forecasting models that aim to predict SEP event occurrence based off of active region configurations in magnetograms (e.g., MAG4, MagPy) have a very challenging task to discern between conditions that will produce flares and conditions that will produce flares and SEPs. However, these are the only types of models that are likely able to provide significant advance warning, if they gain skill. Studies to identify new or supplemental pre-eruptive parameters with predictive power are needed. This may include extracting more information from magnetogram data or taking advantage of signatures in EUV or other types of imaging.

CME-based models have demonstrated skill, but due to delays in receiving the necessary CME parameters, they struggle to produce forecasts on a timeframe useful for operations. Any reduction of time to input CME parameters into the models would improve advance warning. Tools, data availability, or other ways to speed up CME measurements (e.g. high cadence coronagraphs, low latency data streams, automated extraction of CME parameters) would be beneficial.

Accuracy of peak flux forecasts must be increased while scatter is decreased in order for peak flux forecasts to become reliable enough for operational use. This implies that the conditions responsible for event-to-event variability must be better understood and reproduced by models. The paradigm must shift from predicting statistically average outcomes to predicting the outcome for **this** event.

Models should optimize on continuous metrics, as well as categorical (i.e. HSS); for example, reducing mean log error for peak flux models. Categorical metrics treat predictions just above and just below threshold differently even though they are very close in value. Optimizing on a categorical metric effectively puts predictions like peak flux on a coarser scale, potentially reducing accuracy. Using continuous met-

rics will reduce the error between observations and predictions, improving accuracy without assuming specific SEP event definitions or thresholds.

8.3 Model Requirements and Technology Gaps for Radiation Protection

SRAG published requirements for space weather forecasting models in "Space Weather Monitoring and Modeling Requirements for Beyond-LEO Missions" (document no. AES-CHP-SW-002) in Dec 2020. The requirements were based on the understanding of the state of the art at the time, an understanding that is superseded by the analysis presented in this technical report. As a result, SRAG's model requirements need to be re-visited in light of advances in our understanding of model performance and the updated concept of forecasting model use for Artemis missions. In particular, the referenced state of the art for categorical metrics (All Clear) and probabilistic metrics were based on SWPC performance as analyzed in Bain et al. (2021), which as we showed in Section 6.1 includes "hits" due to persistence forecasts issued during an ongoing event. Such forecasts are not reflective of the task of predicting future onsets, and their inclusion significantly boosts the metrics to values that are practically unattainable at this point of the technology development (e.g. False Alarm Ratio of 0.16, TSS=0.61, HSS=0.70, see also Table 6.3).

Current versions of technology gap documents pertaining to space weather forecasting also make reference to performance metrics considered the state of the art at the time, as well as goals for completing the gaps expressed in terms of multiplicative factors of performance improvement. As above, these should be revisited with consideration to the new understanding of the state of the art presented in this report. Additionally, the use of the HSS as a metric basis to measure advancement of model performance must be done in a careful manner. As discussed in Section 3.4, the imbalance of a validation dataset critically impacts the relative numerical value of this metric: e.g. a $HSS = 0.5$ model in a highly imbalanced dataset is significantly more difficult to achieve than the same score in a balanced dataset. As a result, **gap closure and model requirement tracking should specify the benchmark validation dataset or the measure of imbalance when using the HSS**. The expression of desired improvement in terms of multiplicative factors to the HSS is an unclear goal post, as the base performance improvements (e.g. in Hit Rate, True Negative Rate) needed to achieve a multiple also depend on the imbalance of the dataset. It is worthwhile to consider metrics that are not sensitive to dataset imbalance when looking for incremental improvements in performance, like the TSS. Optimizing model performance according to TSS has been identified as less useful for SRAG operations, however, it is suitable for demonstrating incremental improvements in model performance that is robust to dataset imbalance.

Finally, the development of the SEP Scoreboard and the SPHINX validation framework now makes it possible to rigorously validate models and compare their performance to the stated requirements. Procedures and processes should be developed to use these tools to qualify models for use in operations, using the lessons learned from this validation effort. A more rigorous Research-to-Operations (R2O)

procedure will benefit operators by introducing quality control into the forecasting model pipeline, and will benefit model developers by establishing clear goals for which to focus their efforts.

8.4 Conclusion

The SPHINX validation framework, using forecast data accumulated on the SEP Scoreboard and through the SEPVAL challenge, has established the current state-of-the-art performance in forecasting solar energetic particle events for the purposes of radiation protection for human spaceflight. The validation analysis presented in this report has evaluated 11 models and one national forecasting office in detail, and several more models participating in SEPVAL at a high-level, providing a rich source of insights into model performance among various aspects, from binary All Clear forecasts, probabilistic forecasts, to peak flux and advance warning time. The landscape of SEP forecasting models is complex, and interpreting the results and fairly comparing them requires a high level of understanding of the varied model behaviors and circumstances with respect to the validation dataset. We hope that this work will long serve as a useful resource for understanding forecasting model performance, and will serve as a foundation for NASA and its partners to advance the technology of space weather prediction for the purpose of protecting the next generation of astronauts venturing into deep space from the risks of radiation exposure from solar particle events.

References

A. Ahmadzadeh, D. J. Kempton, P. C. Martens, and R. A. Angryk. Contingency space: A semimetric space for classification evaluation. *IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE*, 45 (2), 2023. doi: https://doi.org/10.1109/TPAMI.2022.3167007. URL https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9756878&tag=1. metrics and validation.

A. Anastasiadis, A. Papaioannou, I. Sandberg, M. Georgoulis, K. Tziotziou, A. Kouloumvakos, and P. Jiggens. Predicting flares and solar energetic particle events: the forspef tool. *Solar Physics*, 292:134, 09 2017. doi: 10.1007/s11207-017-1163-7.

A. Aran, B. Sanahuja, and D. Lario. Solpenco: A solar particle engineering code. *Advances in Space Research*, 37(6):1240–1246, 2006. ISSN 0273-1177. doi: https://doi.org/10.1016/j.asr.2005.09.019. URL https://www.sciencedirect.com/science/article/pii/S0273117705011312. Space weather prediction: Applications and validation.

H. M. Bain, R. A. Steenburgh, T. G. Onsager, and E. M. Stitely. A Summary of National Oceanic and Atmospheric Administration Space Weather Prediction Center Proton Event Forecast Performance and Skill. *Space Weather*, 19(7): e2020SW002670, July 2021. doi: 10.1029/2020SW002670.

D. G. Bonnett and T. A. Wright. Sample Size Requirements for Estimating Pearson, Kendall and Spearmann Correlations. *Psychometrika*, 65(1):23–28, Mar. 2000. doi: 0033-3123/2000-1/1997-0607-A.

A. Bruno. Calibration of the GOES 13/15 high-energy proton detectors based on the PAMELA solar energetic particle observations. *Space Weather*, 15(9):1191–1202, Sept. 2017. doi: 10.1002/2017SW001672.

D. A. Clarke. A Consistent Method of Characteristics for Multidimensional Magnetohydrodynamics. *The Astrophysical Journal*, 457:291, Jan. 1996. doi: 10.1086/176730.

Collaboration for Australian Weather and Climate Research. "wwrp/wgne joint working group on forecast verification research", Jan. 2015. URL https://www.cawcr.gov.au/projects/verification/#Methods_for_probabilistic_forecasts. Accessed: 2025-04-09.

C. A. Doswell III, R. Davies-Jones, and D. L. Keller. On Summary Measures of Skill in Rare Event Forecasting Based on Contingency Tables. *Weather and Forecasting*, 5:576, July 1990. URL https://www.swpc.noaa.gov/sites/default/files/images/u30/Doswell%20III%2C%20C.A.%2C%20R.P.%20Davies-Jones%2C%20and%20D.L.%20Keller%2C%201990.pdf.

D. Falconer, A. F. Barghouty, I. Khazanov, and R. Moore. A tool for empirical forecasting of major flares, coronal mass ejections, and solar particle events from a proxy of active-region free magnetic energy. *Space Weather*, 9(4):S04003, Apr. 2011. doi: 10.1029/2009SW000537.

D. A. Falconer, R. L. Moore, and G. A. Gary. Correlation of the Coronal Mass Ejection Productivity of Solar Active Regions with Measures of Their Global Nonpotentiality from Vector Magnetograms: Baseline Results. *Astrophysical Journal*, 569(2):1016–1025, Apr. 2002. doi: 10.1086/339161.

D. A. Falconer, R. L. Moore, A. F. Barghouty, and I. Khazanov. Prior Flaring as a Complement to Free Magnetic Energy for Forecasting Solar Eruptions. *The Astrophysical Journal*, 757(1):32, Sept. 2012. doi: 10.1088/0004-637X/757/1/32.

D. A. Falconer, R. L. Moore, A. F. Barghouty, and I. Khazanov. MAG4 versus alternative techniques for forecasting active region flare productivity. *Space Weather*, 12(5):306–317, May 2014. doi: 10.1002/2013SW001024.

M. K. Georgoulis and D. M. Rust. Quantitative Forecasting of Major Solar Flares. *Astrophysical Journal Letters*, 661(1):L109–L112, May 2007. doi: 10.1086/518718.

Georgoulis, Manolis K., Bloomfield, D. Shaun, Piana, Michele, Massone, Anna Maria, Soldati, Marco, Gallagher, Peter T., Pariat, Etienne, Vilmer, Nicole, Buchlin, Eric, Baudin, Frederic, Csillaghy, Andre, Sathiapal, Hanna, Jackson, David R., Alingery, Pablo, Benvenuto, Federico, Campi, Cristina, Florios, Konstantinos, Gontikakis, Constantinos, Guennou, Chloe, Guerra, Jordan A., Kontogiannis, Ioannis, Latorre, Vittorio, Murray, Sophie A., Park, Sung-Hong, von Stachelski, Samuel, Torbica, Aleksandar, Vischi, Dario, and Worsfold, Mark. The flare likelihood and region eruption forecasting (flarecast) project: flare forecasting in the big data & machine learning era. *J. Space Weather Space Clim.*, 11:39, 2021. doi: 10.1051/swsc/2021023. URL https://doi.org/10.1051/swsc/2021023.

J. Hu, G. Li, X. Ao, G. P. Zank, and O. Verkhoglyadova. Modeling particle acceleration and transport at a 2-d cme-driven shock. *Journal of Geophysical Research: Space Physics*, 122(11):10,938–10,963, 2017. doi: https://doi.org/10.1002/2017JA024077. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2017JA024077.

S. Hu, M.-H. Y. Kim, G. E. McClellan, and F. Cucinotta. Modeling the acute health effects of astronauts from exposure to large solar particle events. *Health Physics*, 96(4):465–476, 2009. doi: 10.1097/01.HP.0000339020.92837.61.

S. W. Kahler and A. G. Ling. Forecasting Solar Energetic Particle (SEP) events with Flare X-ray peak ratios. *Journal of Space Weather and Space Climate*, 8: A47, Oct. 2018. doi: 10.1051/swsc/2018033.

M. W. Liemohn, A. D. Shane, A. R. Azari, A. K. Petersen, B. M. Swiger, and A. Mukhopadhyay. Rmse is not enough: Guidelines to robust data-model comparisons for magnetospheric physics. *Journal of Atmospheric and Solar-Terrestrial*

Physics, 218:105624, 2021. ISSN 1364-6826. doi: https://doi.org/10.1016/j.jastp.2021.105624. URL https://www.sciencedirect.com/science/article/pii/S1364682621000857.

J. G. Luhmann, S. A. Ledvina, D. Krauss-Varban, D. Odstrcil, and P. Riley. A heliospheric simulation-based approach to SEP source and transport modeling. *Advances in Space Research*, 40(3):295–303, Jan. 2007. doi: 10.1016/j.asr.2007.03.089.

J. G. Luhmann, S. A. Ledvina, D. Odstrcil, M. J. Owens, X. P. Zhao, Y. Liu, and P. Riley. Cone model-based SEP event calculations for applications to multipoint observations. *Advances in Space Research*, 46(1):1–21, July 2010. doi: 10.1016/j.asr.2010.03.011.

M. Núñez. Predicting solar energetic proton events (E > 10 MeV). *Space Weather*, 9(7):S07003, July 2011. doi: 10.1029/2010SW000640.

M. Núñez. Real-time prediction of the occurrence and intensity of the first hours of >100 MeV solar energetic proton events. *Space Weather*, 13(11):807–819, Nov. 2015. doi: 10.1002/2015SW001256.

M. Núñez and D. Paul-Pena. Predicting >10 MeV SEP Events from Solar Flare and Radio Burst Data. *Universe*, 6(10):161, Sept. 2020. doi: 10.3390/universe6100161.

M. Núñez, P. J. Reyes-Santiago, and O. E. Malandraki. Real-time prediction of the occurrence of GLE events. *Space Weather*, 15(7):861–873, July 2017. doi: 10.1002/2017SW001605.

E. Palmerio, C. O. Lee, M. L. Mays, J. G. Luhmann, D. Lario, B. Sánchez-Cano, I. G. Richardson, R. Vainio, M. L. Stevens, C. M. S. Cohen, K. Steinvall, C. Möstl, A. J. Weiss, T. Nieves-Chinchilla, Y. Li, D. E. Larson, D. Heyner, S. D. Bale, A. B. Galvin, M. Holmström, Y. V. Khotyaintsev, M. Maksimovic, and I. G. Mitrofanov. CMEs and SEPs During November-December 2020: A Challenge for Real-Time Space Weather Forecasting. *Space Weather*, 20(5):e2021SW002993, May 2022. doi: 10.1029/2021SW002993.

A. Papaioannou, A. Anastasiadis, I. Sandberg, M. K. Georgoulis, G. Tsiropoula, K. Tziotziou, P. Jiggens, and A. Hilgers. A Novel Forecasting System for Solar Particle Events and Flares (FORSPEF). In *Journal of Physics Conference Series*, volume 632 of *Journal of Physics Conference Series*, page 012075. IOP, Aug. 2015. doi: 10.1088/1742-6596/632/1/012075.

A. Papaioannou, R. Vainio, O. Raukunen, P. Jiggens, A. Aran, M. Dierckxsens, S. A. Mallios, M. Paassilta, and A. Anastasiadis. The probabilistic solar particle event forecasting (PROSPER) model. *Journal of Space Weather and Space Climate*, 12:24, May 2022. doi: 10.1051/swsc/2022019.

A. Papaioannou, G. Vasalos, K. Whitman, P. Quinn, A. Anastasiadis, M. Mays, J. Barzilla, C. Didigu, C. Light, C. Corti, J. Jones, A. Chulaki, H. Hermann,

and E. Semones. Exploring the Validation Results of the Advanced Solar Particle Events Casting System (ASPECS). *Space Weather*, 10 2025.

Papaioannou, Athanasios, Sandberg, Ingmar, Anastasiadis, Anastasios, Kouloumvakos, Athanasios, Georgoulis, Manolis K., Tziotziou, Kostas, Tsiropoula, Georgia, Jiggens, Piers, and Hilgers, Alain. Solar flares, coronal mass ejections and solar energetic particle event characteristics. *J. Space Weather Space Clim.*, 6: A42, 2016. doi: 10.1051/swsc/2016035. URL https://doi.org/10.1051/swsc/2016035.

A. Posner. Up to 1-hour forecasting of radiation hazards from solar energetic ion events with relativistic electrons. *Space Weather*, 5(5):05001, May 2007. doi: 10.1029/2006SW000268.

I. G. Richardson, T. T. von Rosenvinge, H. V. Cane, E. R. Christian, C. M. S. Cohen, A. W. Labrador, R. A. Leske, R. A. Mewaldt, M. E. Wiedenbeck, and E. C. Stone. $>$ 25 MeV Proton Events Observed by the High Energy Telescopes on the STEREO A and B Spacecraft and/or at Earth During the First $\sim$ Seven Years of the STEREO Mission. *Solar Physics*, 289(8):3059–3107, Aug. 2014. doi: 10.1007/s11207-014-0524-8.

I. G. Richardson, M. L. Mays, and B. J. Thompson. Prediction of Solar Energetic Particle Event Peak Proton Intensity Using a Simple Algorithm Based on CME Speed and Direction and Observations of Associated Solar Phenomena. *Space Weather*, 16(11):1862–1881, Nov. 2018. doi: 10.1029/2018SW002032.

I. Sandberg, P. Jiggens, D. Heynderickx, and I. A. Daglis. Cross calibration of NOAA GOES solar proton detectors using corrected NASA IMP-8/GME data. *Geophysical Research Letters*, 41(13):4435–4441, July 2014. doi: 10.1002/2014GL060469.

D. F. Smart and M. A. Shea. Comment on the use of GOES solar proton data and spectra in solar proton dose calculations. *Radiation Measurements*, 30(3): 327–335, June 1999. doi: 10.1016/S1350-4487(99)00059-1.

C. Spearman. The proof and measurement of association between two things. *The American Journal of Psychology*, 15(1):72–101, 1904. ISSN 00029556. URL http://www.jstor.org/stable/1412159.

E. Weisstein. "statistical correlation", 2025. URL https://mathworld.wolfram.com/StatisticalCorrelation.html. Accessed: 2025-04-09.

K. Whitman, R. Egeland, I. G. Richardson, C. Allison, P. Quinn, J. Barzilla, I. Kitiashvili, V. Sadykov, H. M. Bain, M. Dierckxsens, M. L. Mays, T. Tadesse, K. T. Lee, E. Semones, J. G. Luhmann, M. Núñez, S. M. White, S. W. Kahler, A. G. Ling, D. F. Smart, M. A. Shea, V. Tenishev, S. F. Boubrahimi, B. Aydin, P. Martens, R. Angryk, M. S. Marsh, S. Dalla, N. Crosby, N. A. Schwadron, K. Kozarev, M. Gorby, M. A. Young, M. Laurenza, E. W. Cliver, T. Alberti, M. Stumpo, S. Benella, A. Papaioannou, A. Anastasiadis, I. Sandberg, M. K.

Georgoulis, A. Ji, D. Kempton, C. Pandey, G. Li, J. Hu, G. P. Zank, E. Lavasa, G. Giannopoulos, D. Falconer, Y. Kadadi, I. Fernandes, M. A. Dayeh, A. Muñoz-Jaramillo, S. Chatterjee, K. D. Moreland, I. V. Sokolov, I. I. Roussev, A. Taktakishvili, F. Effenberger, T. Gombosi, Z. Huang, L. Zhao, N. Wijsen, A. Aran, S. Poedts, A. Kouloumvakos, M. Paassilta, R. Vainio, A. Belov, E. A. Eroshenko, M. A. Abunina, A. A. Abunin, C. C. Balch, O. Malandraki, M. Karavolos, B. Heber, J. Labrenz, P. Kühl, A. G. Kosovichev, V. Oria, G. M. Nita, E. Illarionov, P. M. O'Keefe, Y. Jiang, S. H. Fereira, A. Ali, E. Paouris, S. Aminalragia-Giamini, P. Jiggens, M. Jin, C. O. Lee, E. Palmerio, A. Bruno, S. Kasapis, X. Wang, Y. Chen, B. Sanahuja, D. Lario, C. Jacobs, D. T. Strauss, R. Steyn, J. van den Berg, B. Swalwell, C. Waterfall, M. Nedal, R. Miteva, M. Dechev, P. Zucca, A. Engell, B. Maze, H. Farmer, T. Kerber, B. Barnett, J. Loomis, N. Grey, B. J. Thompson, J. A. Linker, R. M. Caplan, C. Downs, T. Török, R. Lionello, V. Titov, M. Zhang, and P. Hosseinzadeh. Review of Solar Energetic Particle Prediction Models. *Advances in Space Research*, 72(12):5161–5242, Dec. 2023. doi: 10.1016/j.asr.2022.08.006.

P. Zucca, M. Núñez, and K. Klein. Exploring the potential of microwave diagnostics in SEP forecasting: The occurrence of SEP events. *Journal of Space Weather and Space Climate*, 7:A13, June 2017. doi: 10.1051/swsc/2017011.

Appendix A

Metrics Produced by SPHINX

SPHINX calculates a wide variety of metrics. Selected validation metrics found to be most informative are described here. These metrics fall into the categories of All Clear ratios and skill scores, probability metrics, flux metrics, and timing metrics. Note that much of the material in this Appendix is modified from Collaboration for Australian Weather and Climate Research (2015).

A.1 All Clear Metrics

All Clear metrics are derived from the components of the contingency table,

	Observed		
	Yes	No	Sum
Pred. Yes	h	f	$h+f$
Pred. No	m	c	$m+c$
Sum	$h+m$	$f+c$	N

where, in this context,

- h (hits) represents the number of events forecasted to occur and actually occurred,
- m (misses) represents the number of events forecasted to *not* occur and actually occurred,
- f (false alarms) represents the number of events forecasted to occur and *did not* actually occur,
- c (correct negatives) represents the number of events forecasted to *not* occur and *did not* actually occur, and
- N is the total number of forecasts.

The All Clear metrics that SPHINX calculates are derived from these integers. Each metric will be described and defined mathematically and then summarized in Table A.1.

Balanced Accuracy: Answers the question: "If correctly forecasting both 'yes' and 'no' events are equally important, how well does the forecasting model perform, on average, at predicting each condition?"

$$BA = \frac{h}{2(h+m)} + \frac{c}{2(c+f)} \tag{A.1}$$

Bias: Answers the question: "How did the forecast frequency of 'yes' events compare to the observed frequency of 'yes' events?"

$$B = \frac{h+f}{h+m} \tag{A.2}$$

Detection Failure Ratio: Answers the question: "What fraction of the 'no' predictions failed to predict an observed 'yes' event?"

$$DFR = \frac{m}{m+c} \tag{A.3}$$

F-Score: Answers the question: "If one views recall as β times as important as precision for a forecast, how well does the forecasting model perform?"

$$\mathcal{F}(\beta) = \frac{h(1+\beta^2)}{h(1+\beta^2)+f+m\beta^2} \tag{A.4}$$

False Alarm Rate: Answers the question: "What fraction of the observed 'no' events were incorrectly forecast as 'yes'?" Also called **False Positive Rate**.

$$F = \frac{f}{f+c} \tag{A.5}$$

False Alarm Ratio: Answers the question: "What fraction of the predicted 'yes' events actually did not occur (i.e. were False Alarms)?" Also called **False Discovery Rate**.

$$FAR = \frac{f}{h+f} \tag{A.6}$$

False Alarm Event Ratio: Answers the question: "How many false alarms were issued compared to observed events?" The acronym FAER is pronounced "fear". FAER ranges from 0 to infinity with a desired value of 1 or less. This metric is proposed here to guide models towards performance that is reliable enough for use in operations and acts as a constraint on the False Alarm Rate when forecasting in a true climatologically imbalanced scenario.

$$F = \frac{f}{h+m} \tag{A.7}$$

Fowlkes-Mallows Index: Answers the question: "How well does the forecasting model balance correctly predicting 'yes' events (without too many False Alarms) and successfully detecting most 'yes' events when they occur?"

$$FMI = \sqrt{\frac{h}{h+f}\frac{h}{h+m}} \tag{A.8}$$

Frequency of Correct Negatives: Answers the question: "How often did the forecasting model correctly predict 'no' events?" Also called **Negative Predictive Value**.

$$FCN = \frac{c}{c+m} \tag{A.9}$$

Frequency of Hits: Answers the question: "How often did the forecasting model correctly predict 'yes' events?" Also called **Precision** or **Positive Predictive Value**.

$$FH = \frac{h}{h+f} \tag{A.10}$$

Frequency of Misses: Answers the question: "How often did the forecasting model fail to predict 'yes' events?" Also called **False Negative Rate**.

$$FM = \frac{m}{m+h} \tag{A.11}$$

Gilbert Skill Score: Answers the question: "How skillfully does the forecasting model predict 'yes' events beyond what would be expected by randomly guessing?"

$$GSS = \frac{h - \frac{(h+f)(h+m)}{N}}{h+f+m-\frac{(h+f)(h+m)}{N}} \tag{A.12}$$

Heidke Skill Score: Answers the question: "What was the accuracy of the forecast relative to that of random chance?"

$$HSS = \frac{2(hc-fm)}{(h+m)(m+c)+(h+f)(f+c)} \tag{A.13}$$

Hit Rate: Answers the question: "What fraction of the observed 'yes' events were correctly forecast?" Also called **Recall** and **Sensitivity**.

$$H = \frac{h}{h+m} \tag{A.14}$$

Informedness: Answers the question: "How well does the forecasting model detect reality?"

$$I = \frac{h}{h+m} + \frac{c}{c+f} - 1 \tag{A.15}$$

Markedness: Answers the question: "How reliable are the forecasting model's predictions?"

$$M = \frac{h}{h+f} + \frac{c}{c+m} - 1 \tag{A.16}$$

Matthew Correlation Coefficient: Answers the question: "Does the forecasting model make predictions that are reliably consistent with reality across 'yes' and 'no' event conditions, even if 'yes' events are rare and 'no' events are not?"

$$MCC = \frac{hc-fm}{\sqrt{(h+f)(h+m)(c+f)(c+m)}} \tag{A.17}$$

Odds Ratio: Answers the question: "What is the ratio of the odds of a 'yes' forecast being correct, to the odds of a 'yes' forecast being wrong?"

$$OR = \frac{hc}{mf} \tag{A.18}$$

Odds Ratio Skill Score: Answers the question: "How much better are this forecasting model's odds of correctly predicting a 'yes' or 'no' event than a random guess?"

$$ORSS = \frac{hc-mf}{hc+mf} \tag{A.19}$$

Percent Correct: Answers the question: "What is the fraction of correct forecasts?" Also called **Accuracy**.

$$PC = \frac{h + c}{N} \tag{A.20}$$

Prevalence: Answers the question: "What is the fraction of 'yes' events compared to the total number of 'yes' and 'no' events?"

$$P = \frac{h + m}{N} \tag{A.21}$$

Prevalence Threshold: This is less a performance metric and more of a statistic that indicates the prevalence level at which the false positive rate increases most rapidly.

$$PT = \frac{\sqrt{\frac{h}{h+m}\frac{f}{f+c}} - \frac{f}{f+c}}{\frac{h}{h+m} - \frac{f}{f+c}} \tag{A.22}$$

Probability of Correct Negatives: Answers the question "What is the probability that the forecasting model correctly forecasts observed 'no' events?" Also called **Specificity**, **Selectivity**, **True Negative Rate**.

$$PCN = \frac{c}{c + f} \tag{A.23}$$

Symmetric Extreme Dependency Score: Answers the question: "How does the forecasting model perform for rare events?"

$$SEDS = \frac{\ln\left(\frac{h+f}{N}\right) + \ln\left(\frac{h+m}{N}\right)}{\ln\left(\frac{h}{N}\right)} \tag{A.24}$$

Threat Score: Answers the question: "How well did the forecast 'yes' events correspond to the observed 'yes' events?"

$$TS = \frac{h}{h + f + m} \tag{A.25}$$

True Skill Score: Answers the question: "How well did the forecast separate the 'yes' events from the 'no' events?"

$$TSS = H - F = \frac{h}{h + m} - \frac{f}{f + c} \tag{A.26}$$

Name	Attribute	Equation	Range	Perfect Score
Balanced Accuracy	Accuracy	A.1	0 to 1	1
Bias	Bias	A.2	0 to ∞	1
Detection Failure Ratio	Resolution	A.3	0 to 1	0
F-Score	Classification	A.4	0 to 1	1
False Alarm Rate	Discrimination	A.5	0 to 1	0
False Alarm Ratio	Resolution	A.6	0 to 1	1
Fowlkes-Mallows Index	Classification/Clustering	A.8	0 to 1	1
Frequency of Correct Negatives	Resolution	A.9	0 to 1	1
Frequency of Hits	Resolution	A.10	0 to 1	1
Frequency of Misses	Discrimination	A.11	0 to 1	0
Gilbert Skill Score		A.12	$-\infty$ to 1	1
Heidke Skill Score		A.13	$-\infty$ to 1	1
Hit Rate	Discrimination	A.14	0 to 1	1
Informedness		A.15	0 to 1	1
Markedness		A.16	0 to 1	1
Matthew Correlation Coefficient	Association	A.17	-1 to 1	1
Odds Ratio	Accuracy	A.18	0 to ∞	∞
Odds Ratio Skill Score		A.19	-1 to 1	1
Percent Correct	Accuracy	A.20	0 to 1	1
Prevalence	Event List Statistic	A.21	0 to 1	1
Prevalence Threshold	Percentage	A.22	0 to ∞	
Probability of Correct Negatives	Discrimination	A.23	0 to 1	1
Symmetric Extreme Dependence Score		A.24	$-\infty$ to 1	1
Threat Score	Accuracy/Clustering	A.25	0 to 1	1
True Skill Score		A.26	-1 to 1	1

Table A.1: All Clear metrics calculated in SPHINX.

A.2 Probability Metrics

Brier Score: Answers the question: "How large was the error in the forecast?"

$$BS = \frac{1}{N}\sum_{i=1}^{N}(p_i - o_i)^2, \tag{A.27}$$

where p_i is the forecasted probability of occurrence (of an SEP event) for forecast i, and o_i is the observed probability of occurrence (e.g., 0 or 1) for forecast i.

Brier Skill Score: Answers the question: "What is the relative skill of the probabilistic forecast over climatology, in terms of predicting whether or not an event occurred?"

$$BSS = 1 - \frac{BS}{BS_{ref}} \tag{A.28}$$

In SPHINX, $BS_{ref} = 0.033$, according to the climatological SEP probability value given in Bain et al. (2021).

Area Under ROC Curve: Answers the question: "On average, as a function of False Alarm Rate, how much better or worse is this forecasting model compared to a random forecast?"

$$AUC = \int_0^1 H(FAR)\mathrm{d}FAR, \tag{A.29}$$

where $H(FAR)$ is the Hit Rate as a function of False Alarm Rate. $AUC = 0.5$ represents a random forecast. $AUC > 0.5$ represents a forecast that is better than random. $AUC < 0.5$ represents a forecast that is worse than random.

Spearman (Rank) Correlation Coefficient: Answers the question: "How well can the relationship between predicted SEP probability and observed SEP probability be described using a monotonic function?" Compare to the Pearson Correlation Coefficient, which represents the same thing, but limited to *linear* correlations.

$$SCC = \frac{\mathrm{cov}(R(x), R(y))}{\sigma_{R(x)}\sigma_{R(y)}}, \tag{A.30}$$

where x and y represent vectors of predicted and observed SEP probabilities, $R(\cdot)$ is the rank function—which reassigns the elements of the argument vector and replaces them with their rank integer—and $\sigma_R(\cdot)$ is the standard deviation of the rank data.

Name	Attribute	Equation	Range	Perfect Score
Brier Score		$BS = \frac{1}{N}\sum_{i=1}^{N}(p_i - o_i)^2$	0 to 1	1
Brier Skill Score		$BSS = 1 - \frac{BS}{BS_{ref}}$	-∞ to 1	1
Area Under ROC Curve		$AUC = \int_0^1 H(FAR)\mathrm{d}FAR$	0 to 1	1
Spearman (Rank) Correlation Coefficient		$SCC = \frac{\mathrm{cov}(R(x),R(y))}{\sigma_{R(x)}\sigma_{R(y)}}$	-1 to 1	1

Table A.2: Probability metrics calculated in SPHINX.

A.3 Flux Metrics

Error: Answers the question: "What is the difference between the predicted and observed flux?" Measure of bias. A positive value indicates the model overpredicts while a negative value indicates the model underpredicts.

$$E = f_{\text{pred}} - f_{\text{obs}} \tag{A.31}$$

Absolute Error: Answers the question: "What is the absolute difference between the predicted and observed flux?" Measure of accuracy.

$$AE = |f_{\text{pred}} - f_{\text{obs}}| \tag{A.32}$$

Log Error: Answers the question: "What is the difference between the log of the predicted flux and log of the observed flux?" Measure of bias. A positive value indicates the model overpredicts while a negative value indicates the model underpredicts.

$$LE = \log_{10}(f_{\text{pred}}) - \log_{10}(f_{\text{obs}}) \tag{A.33}$$

Absolute Log Error: Answers the question: "What is the absolute difference between the log of the predicted flux and log of the observed flux?" Measure of accuracy.

$$ALE = |\log_{10}(f_{\text{pred}}) - \log_{10}(f_{\text{obs}})| \tag{A.34}$$

Squared Error: Answers the question: "What is the squared difference between the predicted and observed flux?" Measure of bias. Larger values indicate overprediction/underprediction. More sensitive to outliers than error.

$$SE = (f_{\text{pred}} - f_{\text{obs}})^2 \tag{A.35}$$

Squared Log Error: Answers the question: "What is the squared difference between the log of the predicted flux and log of the observed flux?" Measure of bias. Larger values indicate overprediction/underprediction. More sensitive to outliers than log error.

$$SLE = (\log_{10}(f_{\text{pred}}) - \log_{10}(f_{\text{obs}}))^2 \tag{A.36}$$

Root Mean Squared Error: Answers the question: "On average, what is the error between the predicted flux and the observed flux?"

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(f_{\text{pred},i} - f_{\text{obs},i})^2} \tag{A.37}$$

Root Mean Squared Log Error: Answers the question: "On average, what is the error between the log of the predicted flux and the log of the observed flux?"

$$RMSLE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\log_{10}(f_{\text{pred},i}) - \log_{10}(f_{\text{obs},i}))^2} \tag{A.38}$$

Relative Error: Answers the question: "How significant is the difference between the predicted flux and the observed flux relative to the observed flux?"

$$RE = \frac{f_{\text{pred}} - f_{\text{obs}}}{f_{\text{obs}}} \tag{A.39}$$

Absolute Relative Error: Answers the question: "How significant is the absolute difference between the predicted flux and the observed flux relative to the observed flux?"

$$ARE = \left| \frac{f_{\text{pred}} - f_{\text{obs}}}{f_{\text{obs}}} \right| \tag{A.40}$$

Symmetric Relative Error: Answers the question: "How significant is the difference between the predicted flux and the observed flux relative to the average predicted and observed flux values?"

$$SRE = \frac{2(f_{\text{pred}} - f_{\text{obs}})}{f_{\text{pred}} - f_{\text{obs}}} \tag{A.41}$$

Symmetric Absolute Relative Error: Answers the question: "How significant is the absolute difference between the predicted flux and the observed flux relative to the average magnitude of predicted and observed flux values?"

$$SARE = \frac{2|f_{\text{pred}} - f_{\text{obs}}|}{|f_{\text{pred}}| - |f_{\text{obs}}|} \tag{A.42}$$

Pearson Correlation Coefficient: Answers the question: "How strongly, and in what direction, are the predicted flux and observed flux linearly related?"

$$PCC = \frac{\text{cov}(x, y)}{\sigma_x \sigma_y}, \tag{A.43}$$

where x and y represent vectors of predicted and observed flux values, and σ_x represents the standard deviation of the data contained in the vector x.

Mean Accuracy Ratio: Answers the question: "On average, how close are the predicted fluxes compared to the observed fluxes expressed as a proportion of the observed fluxes?"

$$MAR = \frac{1}{N} \sum_{i=1}^{N} \frac{f_{\text{pred},i}}{f_{\text{obs},i}} \tag{A.44}$$

Median Symmetric Accuracy: Answers the question: "What is the median relative difference between predicted and observed fluxes, treating over- and under-predictions equally?"

$$MedSA = \exp\left(\text{med} \left| \ln(f_{\text{obs}}/f_{\text{pred}}) - 1 \right| \right), \tag{A.45}$$

note that in this equation, f_{obs} and f_{pred} represent vectors and the division between them is element-wise.

Name	Attribute	Equation	Range	Perfect Score
Error	Error	$E = f_{\rm pred} - f_{\rm obs}$	$-\infty$ to ∞	0
Absolute Error	Error	$AE = \lvert f_{\rm pred} - f_{\rm obs} \rvert$	0 to ∞	0
Log Error	Error	$LE = \log_{10}(f_{\rm pred}) - \log_{10}(f_{\rm obs})$	$-\infty$ to ∞	0
Absolute Log Error	Error	$ALE = \lvert \log_{10}(f_{\rm pred}) - \log_{10}(f_{\rm obs}) \rvert$	0 to ∞	0
Relative Error	Error	$RE = \frac{f_{\rm pred} - f_{\rm obs}}{f_{\rm obs}}$	$-\infty$ to ∞	0
Absolute Relative Error	Error	$ARE = \left\lvert \frac{f_{\rm pred} - f_{\rm obs}}{f_{\rm obs}} \right\rvert$	0 to ∞	0
Squared Error	Error	$SE = (f_{\rm pred} - f_{\rm obs})^2$	0 to ∞	0
Squared Log Error	Error	$SLE = (\log_{10}(f_{\rm pred}) - \log_{10}(f_{\rm obs}))^2$	0 to ∞	0
Root Mean Squared Error	Error	$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(f_{{\rm pred},i} - f_{{\rm obs},i})^2}$	0 to ∞	0
Root Mean Squared Log Error	Error	$RMSLE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\log_{10}(f_{{\rm pred},i}) - \log_{10}(f_{{\rm obs},i}))^2}$	0 to ∞	0
Symmetric Relative Error	Error	$SRE = \frac{2(f_{\rm pred} - f_{\rm obs})}{f_{\rm pred} - f_{\rm obs}}$	$-\infty$ to ∞	0
Symmetric Absolute Relative Error	Error	$SARE = \frac{2\lvert f_{\rm pred} - f_{\rm obs} \rvert}{\lvert f_{\rm pred} \rvert - \lvert f_{\rm obs} \rvert}$	$-\infty$ to ∞	0
Pearson Correlation Coefficient		$PCC = \frac{\mathrm{cov}(x,y)}{\sigma_x \sigma_y}$	-1 to 1	1
Median Symmetric Accuracy	Accuracy	$MedSA = \exp\left(\mathrm{med}\left\lvert \ln(f_{\rm obs}/f_{\rm pred}) - 1 \right\rvert\right)$	0 to ∞	
Mean Accuracy Ratio	Accuracy	$MAR = \frac{1}{N}\sum_{i=1}^{N}\frac{f_{{\rm pred},i}}{f_{{\rm obs},i}}$	0 to ∞	1

Table A.3: Flux metrics calculated in SPHINX.

A.4 Time Metrics

Error: Answers the question: "What is the difference between the predicted time and observed time?" Negative values indicate that the predicted time is earlier than observed. Positive values indicate that the predicted time is later than observed.

$$E = t_{\text{pred}} - t_{\text{obs}} \tag{A.46}$$

Advance Warning Time: Answers the question: "What is the difference between the forecast issue time and the observed SEP threshold crossing time?" This indicates how much time ahead of an SEP event that an operator receives a forecast. A positive time indicates that the forecast was issued prior to the SEP start. A negative time indicates that the forecast was issued after the SEP start. Only forecasts with positive AWT can benefit operations. AWT may be calculated with respect to SEP peak time or other important reference times.

$$AWT = t_{\text{obs}} - t_{\text{issue}} \tag{A.47}$$

Appendix B

Acronyms

ACE

Advanced Composition Explorer. 154, 183, 206

ADEPT

Air Force Dynamic Energetic Particle Tool. 59

AI

Artificial Intelligence. 14

AR

active region. 129, 137

ASPECS

Advanced Solar Particle Events Casting System. 59, 65, 153, 224, 225

AU

Astronomical Unit. 182, 183

AUC

Area Under the Curve. 49, 99, 115, 120, 129, 140, 150, 155, 156, 213

AWT

Advance Warning Time. 24, 53, 58, 66, 95, 110, 111, 115, 124, 204, 209, 225, 232, 233, 235, 255

CCMC

Community Coordinated Modeling Center. 9, 14–16, 22, 23, 25, 51, 57, 125

CDAW

Coordinated Data Analysis Workshop. 170

CME

Coronal Mass Ejection. 10, 11, 16, 19, 21, 24, 25, 30–32, 35, 36, 41, 43–45, 53, 59, 61, 62, 65, 66, 68, 72, 82, 94, 95, 99, 104, 110, 115, 118, 123, 129, 153–156, 169–172, 175, 177, 179, 182, 183, 185, 186, 190, 191, 193, 216, 225, 227

COMESEP

COronal Mass Ejections and Solar Energetic Particles. 59, 60, 84

COSPAR

Committee on Space Research. 15

COSTEP

Comprehensive Suprathermal and Energetic Particle Analyzer. 206

DONKI

Database Of Notifications, Knowledge, Information. 25, 32, 59, 61, 62, 65, 66, 68, 110, 170, 175, 225

DSCOVR

Deep Space Climate Observatory. 182

EPAM

Electron, Proton, and Alpha Monitor. 154, 183, 206

ESP

Energetic Storm Particle. 21, 44, 45, 51, 171, 185, 190, 193

ESPE

Energetic Solar Particle Event. 31, 66, 116, 148, 170, 172, 178, 179, 220, 221

EUV

extreme ultraviolet. 14, 65

FAR

False Alarm Ratio. 47, 171, 224

GLE

Ground-Level Event. 198

GME

Goddard Medium Energy. 177

GOES

Geosynchronous Orbit Earth observing Satellite. 25, 57, 147, 154, 169, 177, 198

GSU

Georgia State University. 65, 95, 99, 110, 143, 145, 213

HESPERIA

High Energy Solar Particle Events foRecastIng and Analysis. 65, 111, 115, 198, 206

HMI

Helioseismic and Magnetic Imager. 153

HR

Hit Rate. 53

HSS

Heidke Skill Score. 48, 53–55, 79, 94, 96, 97, 115, 121, 122, 137, 138, 141, 148, 170, 171, 184, 191, 196, 199, 207, 225, 238

ICME

Interplanetary Coronal Mass Ejection. 51, 190

IMF

Interplanetary Magnetic Field. 177, 198

IMP

Interplanetary Monitoring Platform. 177

iPATH

improved Particle Acceleration and Transport in the Heliosphere. 59, 65, 104, 105, 110, 111, 182–188, 192, 220–225

ISEP

Integrated Solar Energetic Proton Event Alert/Warning System. 9, 14–16, 32, 65

ISS

International Space Station. 14

iSWA

integrated Space Weather Analysis System. 22, 57

ISWAT

International Space Weather Action Team. 15

JSOC

Joint Science Operations Center. 65, 137

JSON

JavaScript Object Notation. 19, 20, 22, 23, 51, 151

LASCO

Large Angle and Spectrometric Coronagraph Experiment. 175, 179, 225

LOS

line-of-sight. 128–131, 137, 215

M-FLAMPA

Multiple Field Line Advection Model for Particle Acceleration. 59

M2M

Moon to Mars Space Weather Analysis Office. 9, 14, 16, 25, 30, 31, 59, 61, 62, 65, 66, 68, 110, 155, 156, 170, 175, 225

MAG4

Magnetogram Forecast. 65, 82, 95, 99, 110, 115, 147

MEMPSEP

Multivariate Ensemble of Models for Probabilistic SEP prediction. 59, 60, 84

MHD

magnetohydrodynamic. 169, 182, 186, 190

ML

Machine Learning. 14

NASA JSC

NASA Johnson Space Center. 14

NOAA

National Oceanic and Atmospheric Administration. 57, 58, 65, 71, 116

PAMELA

Payload for Antimatter Matter Exploration and Light-nuclei Astrophysics. 177

PC

Poorly-Connected. 198

pfu

Particle Flux Unit (cm^{-2} s^{-1} sr^{-1}). 51

POCN

Probability of Correct Negatives. 53

POD

Probability of Detection. 53

PPS

Proton Prediction System. 59, 60, 84

R2O

Research-to-Operations. 238

REleASE

Relativistic Electron Alert System for Exploration. 65, 66, 111, 115, 198, 206, 211, 212

ROC

Receiver Operator Characteristic. 49, 82, 119, 120, 129, 131, 135, 136, 140, 141, 145, 150

RSGA

Report of Solar and Geophysical Activity. 116

SAWS

SEP Advanced Warning System. 59, 65, 153, 224

SDO

Solar Dynamics Observatory. 65, 153

SEP

Solar Energetic Particle. 9–26, 30–39, 41–60, 64–68, 72, 75, 77, 79, 82–84, 89, 94–111, 115–121, 123, 124, 126, 128–132, 134–136, 139–141, 143–148, 150–156, 169–172, 175, 177–179, 182–184, 186, 188, 190, 191, 193, 194, 198–200, 202–204, 206, 207, 209, 212, 216, 220, 221, 224–227, 236–239, 250, 255

SEPMOD

Solar Energetic Particle MODel. 59, 65, 66, 104, 110, 190–194, 220–225

SEPSTER

Solar Energetic Particle STEReo. 44, 45, 59, 65, 66, 97, 104, 110, 115, 169–172, 175, 177–179, 216–219, 224, 225

SEPSTER2D

Solar Energetic Particle STEReo 2D. 44, 45, 66, 177–179, 216–219, 224, 225

SEPVAL

SEP Model Validation Working Meeting. 9–11, 15, 29, 30, 32, 38, 49, 59, 61–64, 80, 81, 83–85, 87, 88, 90–93, 96, 116–124, 128, 129, 138–141, 143, 147–151, 170–172, 175, 177–179, 183–186, 188, 190–196, 202–204, 213–216, 218–225, 239

SHARP

Spaceweather HMI Active Region Patch. 128–131, 137, 143

SHINE

Solar Heliospheric and INterplanetary Environment workshop. 15

SOHO

Solar and Heliospheric Observatory. 25, 169, 175, 179, 206

SPE

Solar Particle Event. 31, 51, 66, 116, 148, 196, 213–215

SPHINX

Solar Particles in the Heliosphere validation INfrastructure for SpWX. 9, 10, 15, 16, 18, 20, 22–26, 28, 29, 39–42, 44–46, 48, 56, 58, 89, 110, 116, 117, 123, 124, 202, 238, 239, 245, 249–251, 254

SPREAdFAST

Solar Particle Radiation Environment Analysis and Forecasting – Acceleration and Scattering Transport. 59

SPRINTS

Space Radiation Intelligence System. 59, 65, 74, 83, 84, 99, 111, 147, 148, 150, 151, 224, 225

SRAG

Space Radiation Analysis Group. 9, 14–17, 26, 28–32, 45, 48, 50–52, 54–56, 58, 59, 65, 95, 116, 143, 183, 186, 206, 225, 238

SRS

Solar Region Summary. 64, 71

STAT

SPE Threat Assessment Tool. 59

STEREO

Solar TERrestrial RElations Observatory. 25, 169, 177

SWPC

Space Weather Prediction Center. 11, 14, 15, 26, 31, 64–67, 70–72, 95, 96, 111, 115–120, 122–124, 204, 213–215, 224, 225, 236, 238

SXR

Soft X-Ray. 198

TNR

True Negative Rate. 53–55

TPR

True Positive Rate. 53–55

TSS

True Skill Statistic. 48, 54, 55, 79, 96, 121, 122, 141, 148, 184, 191, 199, 207, 225, 238

UMASEP

University of Málaga Solar Energetic Particle. 59, 60, 65, 66, 70, 74, 95, 97, 104, 105, 111, 115, 198–200, 202–204, 224, 225

UMASOD

University of Málaga predictor from Solar Data. 198

VIVID

Validation in Visually Interactive Displays. 9, 26, 28, 29

WC

Well-Connected. 198

WSA

Wang-Sheeley-Arge. 190

WSA-Enlil

Wang-Sheeley-Arge Enlil. 66

REPORT DOCUMENTATION PAGE	Form Approved OMB No. 0704–0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
01-01-2026	Technical Publication	

4. TITLE AND SUBTITLE
Validation of Solar Energetic Particle Forecasting Models for Space Radiation Operations with SPHINX and VIVID

5a. CONTRACT NUMBER
NNJ15HK11B

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER
SD

6. AUTHOR(S)
Kathryn Whitman, Ricky Egeland, Clayton Allison, Philip Quinn, Luke Stegeman

5d. PROJECT NUMBER

5e. TASK NUMBER
10449.2.04.04.08.1560

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
NASA Johnson Space Center
Houston, Texas

8. PERFORMING ORGANIZATION REPORT NUMBER
L–

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
National Aeronautics and Space Administration
Washington, DC 20546-0001

10. SPONSOR/MONITOR'S ACRONYM(S)
NASA

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
NASA/TP–2026–20260000463

12. DISTRIBUTION/AVAILABILITY STATEMENT
Unclassified-Unlimited
Subject Category
Availability: NASA STI Program (757) 864-9658

13. SUPPLEMENTARY NOTES
An electronic version can be found at http://ntrs.nasa.gov.

14. ABSTRACT

15. SUBJECT TERMS
space weather, forecasting, validation, space radiation

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			STI Information Desk (email: help@sti.nasa.gov)
U	U	U	UU		19b. TELEPHONE NUMBER *(Include area code)* (757) 864-9658