

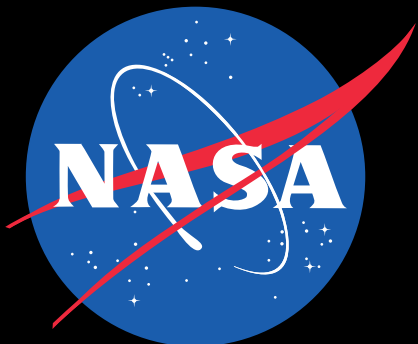
Validation of Solar Energetic Particle Prediction Models for Space Radiation Operations

Kathryn Whitman (KBR), Ricky Egeland (NASA JSC), Clayton Allison (Leidos),
Phil Quinn (Leidos), and Luke Stegeman (Leidos)

NASA JSC Space Radiation Analysis Group

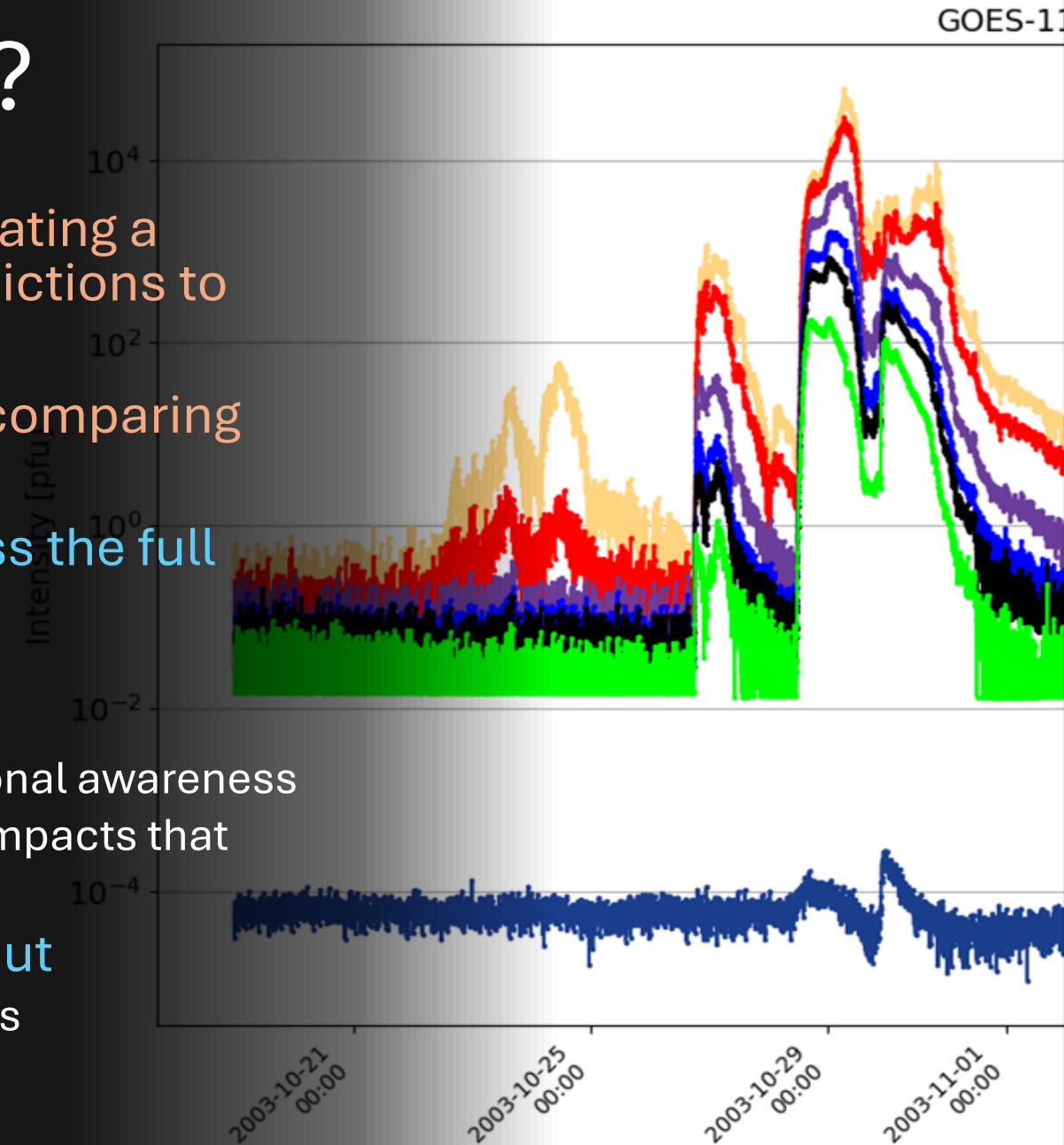
COSPAR Workshop 2026-08-07

Florence, Italy



Why do we do validation?

- I use the term “validation” to mean evaluating a model’s performance by comparing predictions to observations
- Also called “forecast verification” when comparing real-time forecasts to observations
- Understand how a model performs across the full range of observed possibilities
 - No event or typical state of the environment
 - Small events with little or no impact
 - Typical events that require increased situational awareness
 - Large/extreme events with strong or severe impacts that require a response
- Develop intuition to interpret model output
 - Informed use in making operational decisions
 - Informed use for scientific interpretation



SEPVAL Community Validation Effort

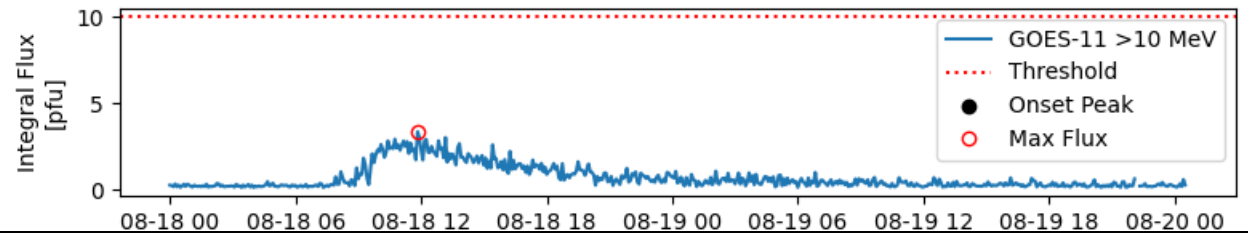
- Ongoing effort since 2019 through SHINE, ISWAT, ESWW
- Historical validation for 63 time periods
 - 33 >10 MeV, 10 pfu SEP events
 - 30 non-event periods
- High-quality historical inputs
 - M2M SWAO remeasured CMEs
- Stronger events that are relevant to space radiation operations
- Balanced dataset (not representative of climatology)
- Same for all models, enabling cross-model comparisons

SEPVAL Dataset

<https://zenodo.org/records/15555244>

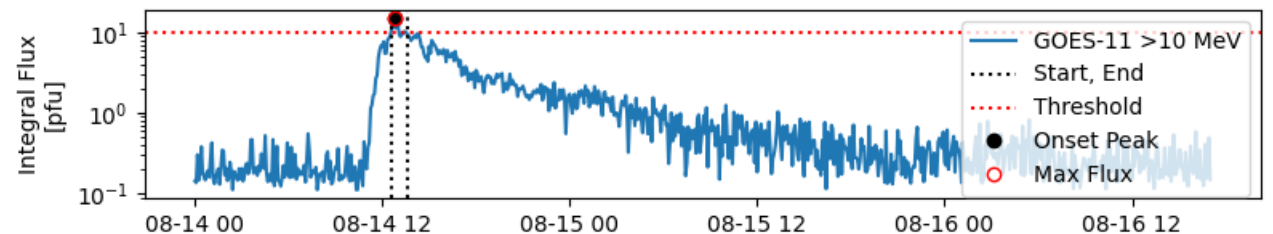
Non-event

Threshold crossings for GOES-11
integral Fluxes



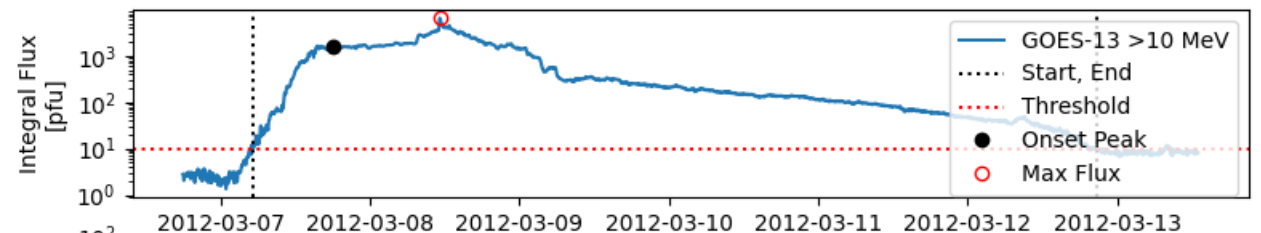
Small event

Threshold crossings for GOES-11
integral Fluxes

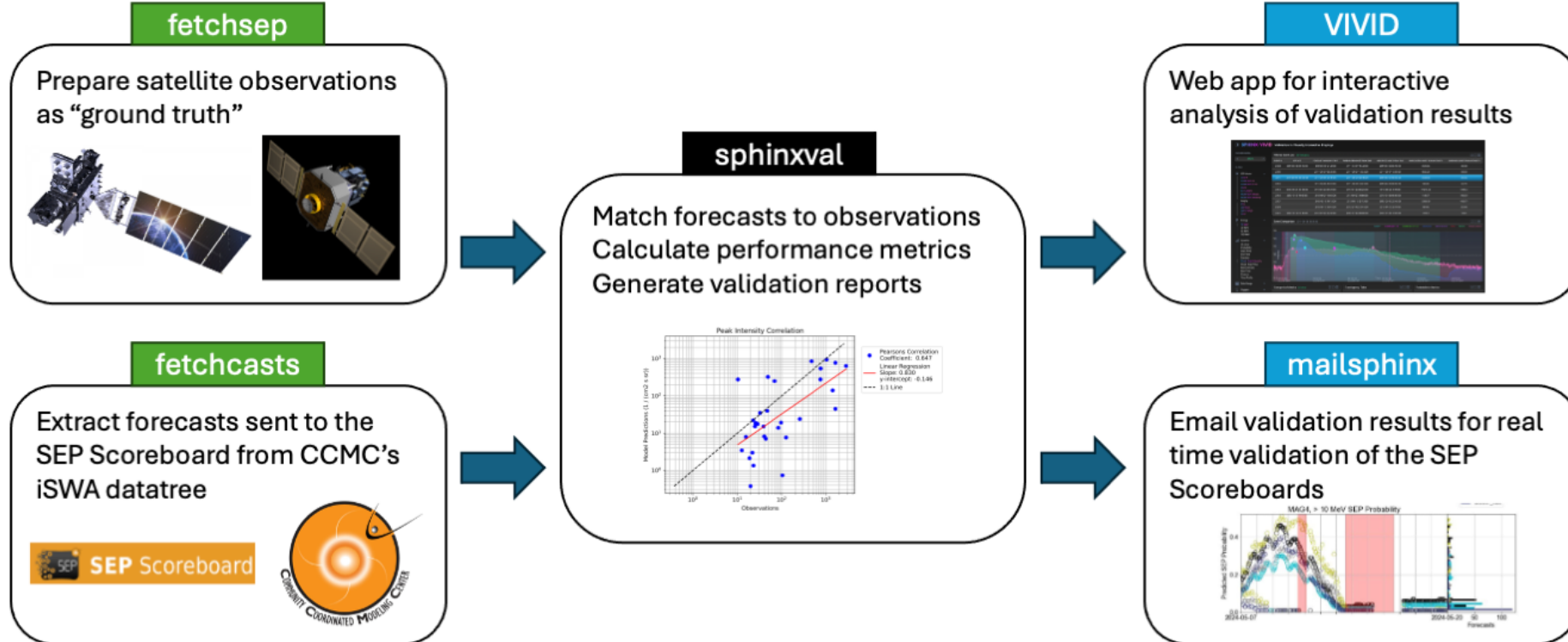


Large event

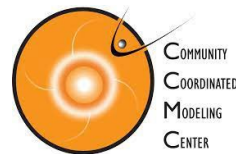
Threshold crossings for GOES-13
integral Fluxes



SPHINX Validation Framework Components



SPHINX is developed at SRAG
in collaboration with:



SEPVAL Participants



SEPVAL Model	Method	Energy Channels (MeV)	Forecast Quantities
ADEPT 1hr, 6hr	Empirical	>10	Time Profile
COMESSEP flare	Empirical	>10	Probability, Peak
COMESSEP 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

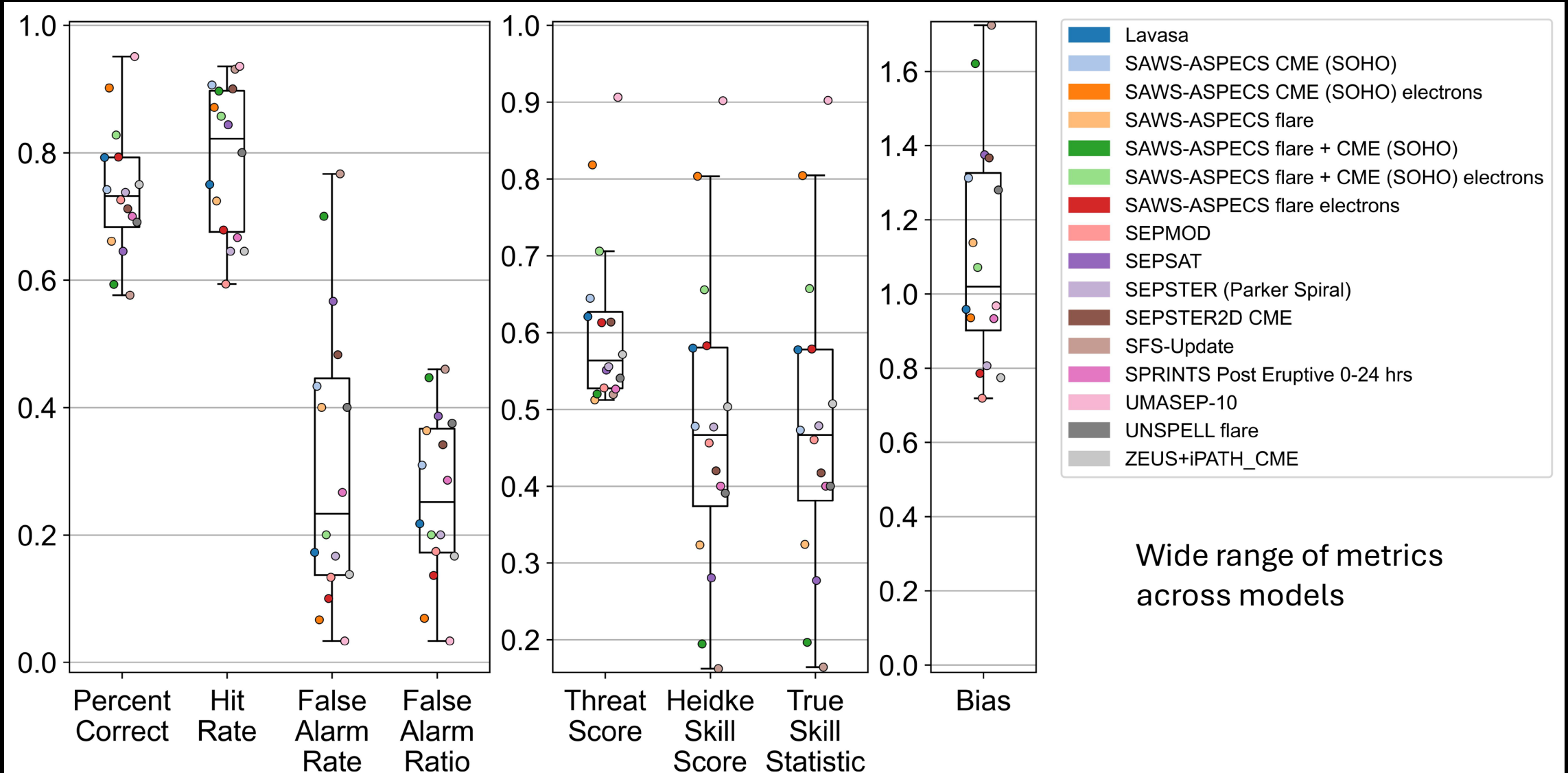
Table 4.3: Model approach, predicted energy channels, and forecasted quantities for models participating in the SEPVAL challenge (in alphabetical order).

Sorted into two groups according to inputs:

- Pre-eruptive  (magnetograms)
- Post-eruptive (flares, CME, particles)
- 23 participating models
- Note that only half the models predict for >100 MeV, a critical energy range for SRAG
- STAT, SPREADFAST, and MFLAMPA were not included in the group metrics that follow

SEPVAL Binary All Clear Results

Post-eruptive Models



SEPVAL All Clear Results

>10 MeV, 10 pfu for Post-Eruptive Models

“Best” depends on choice of metric

Chose to focus on the median value of the group to define state-of-the-art.

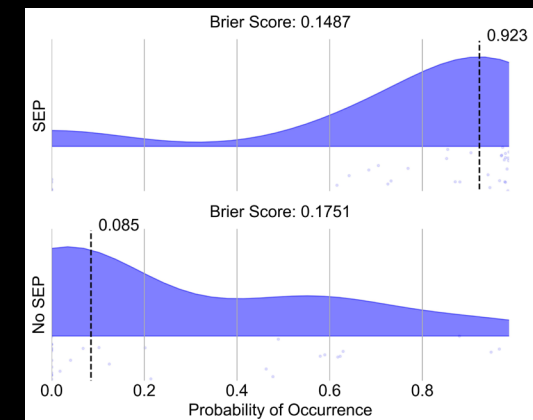
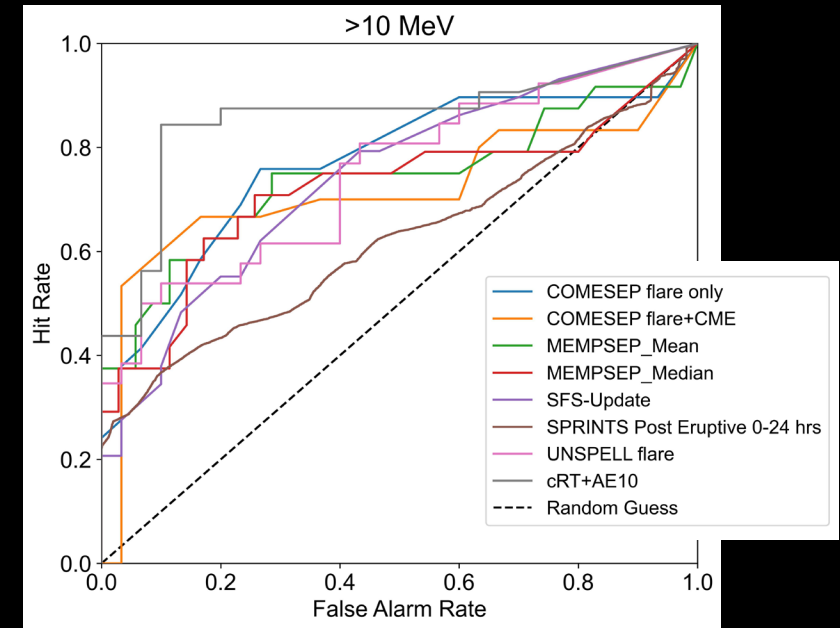
Model	Ranked by Heidke Skill Score
UMASEP-10	0.90
SAWS-ASPECS CME (SOHO) electrons	0.80
SAWS-ASPECS flare + CME (SOHO) electrons	0.66
SAWS-ASPECS flare electrons	0.58
Lavasa	0.58
ZEUS+iPATH_CME	0.50
SAWS-ASPECS CME (SOHO)	0.48
SEPSTER (Parker Spiral)	0.48
ENLIL+SEPMOD	0.46
SEPSTER2D CME	0.42
SPRINTS Post Eruptive 0-24 hrs	0.40
UNSPELL flare	0.39
SAWS-ASPECS flare	0.32
SEPSAT	0.28
SAWS-ASPECS flare + CME (SOHO)	0.19
SFS-Update (PPS Model)	0.16

Model	Ranked by False Alarm Ratio $f/(h+f)$
UMASEP-10	0.03
SAWS-ASPECS CME (SOHO) electrons	0.07
SAWS-ASPECS flare electrons	0.14
ZEUS+iPATH_CME	0.17
ENLIL+SEPMOD	0.17
SAWS-ASPECS flare + CME (SOHO) electrons	0.20
SEPSTER (Parker Spiral)	0.20
Lavasa	0.22
SPRINTS Post Eruptive 0-24 hrs	0.29
SAWS-ASPECS CME (SOHO)	0.31
SEPSTER2D CME	0.34
SAWS-ASPECS flare	0.36
UNSPELL flare	0.38
SEPSAT	0.39
SAWS-ASPECS flare + CME (SOHO)	0.45
SFS-Update (PPS Model)	0.46

SEPVAL Probability of Occurrence

>10 MeV, 10 pfu for Post-Eruptive Models

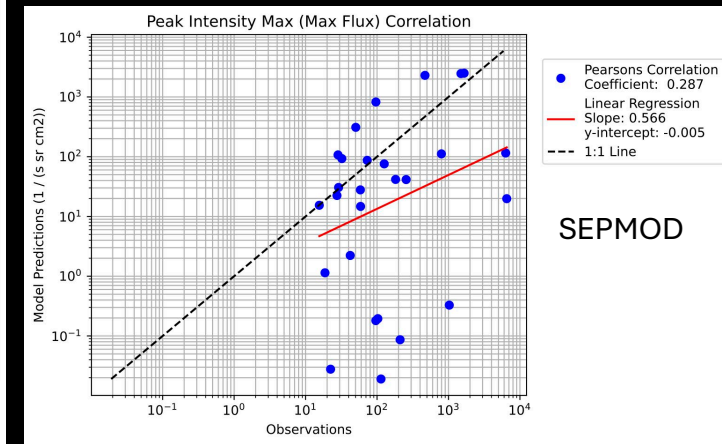
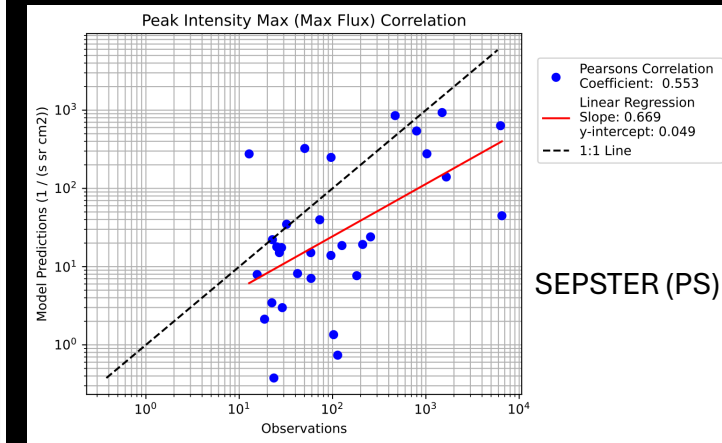
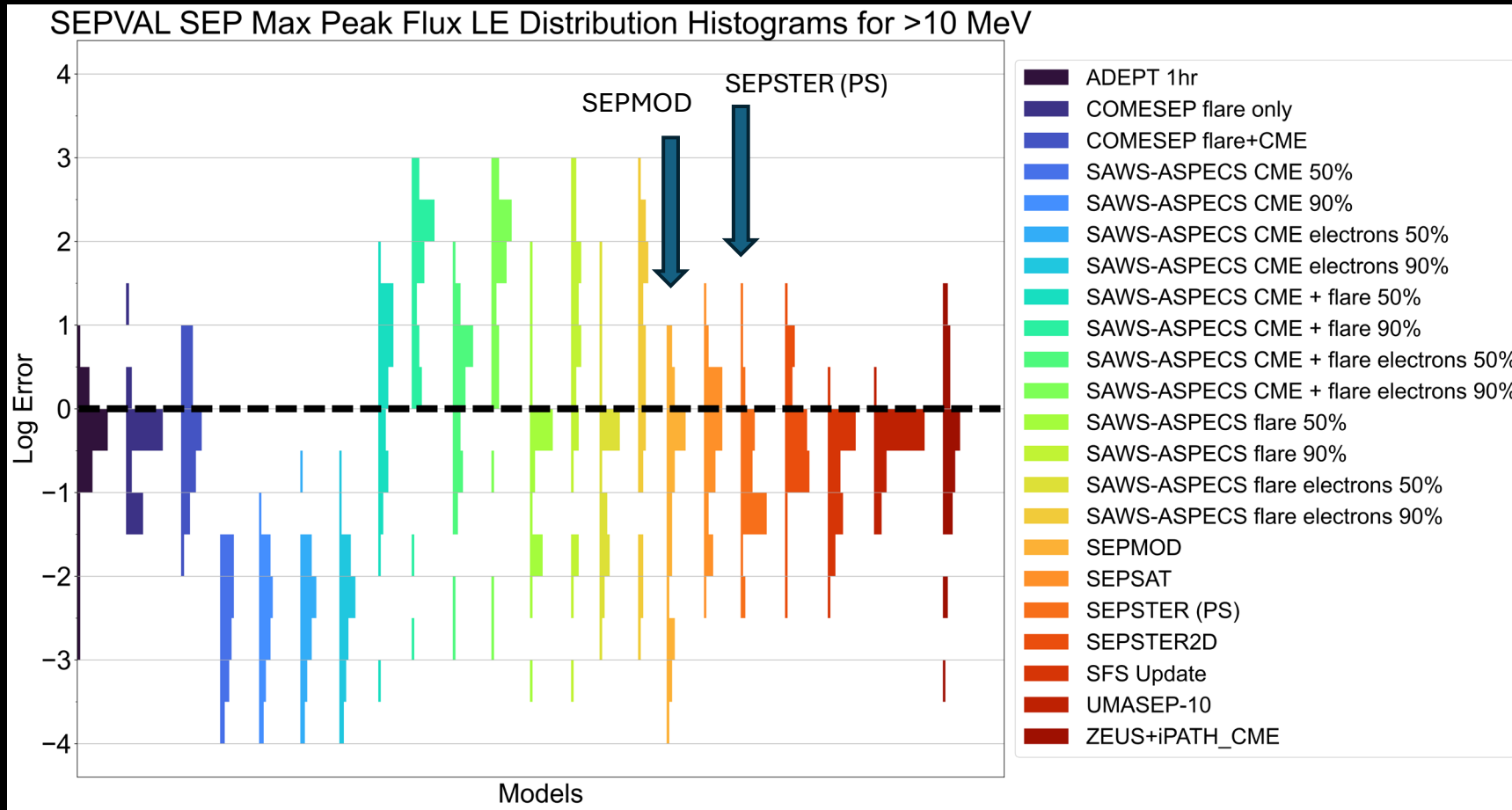
Model	Brier Score	Brier Skill Score	Ranked by Area Under ROC Curve
SAWS-ASPECS CME (SOHO) electrons	0.15	0.69	0.95
SAWS-ASPECS flare + CME (SOHO) electrons	0.11	0.75	0.93
SAWS-ASPECS flare electrons	0.20	0.56	0.90
cRT+AE10	0.16	0.67	0.86
SAWS-ASPECS CME (SOHO)	0.18	0.63	0.84
SAWS-ASPECS flare + CME (SOHO)	0.28	0.40	0.83
COMESep flare only	0.25	0.46	0.77
UNSPELL flare	0.25	0.42	0.75
SFS-Update (PPS Model)	0.26	0.43	0.74
MEMPSEP_Mean	0.20	0.49	0.74
MEMPSEP_Median	0.23	0.40	0.72
COMESep flare+CME	0.23	0.51	0.72
SAWS-ASPECS flare	0.27	0.41	0.69
SPRINTS Post Eruptive 0-24 hrs	0.30	0.34	0.62



cRT_AE10

SEPVAL Maximum Flux Results

>10 MeV, 10 pfu for Post-Eruptive Models



Typical for peak flux to differ by more than an order of magnitude from observations. The large spread in performance from event to event makes it very challenging to use model predictions for decision-making.

SEPVAL Definition of State-of-the-Art

Historical, Balanced Dataset

SEPVAL Pre-eruptive Models				
Metrics	>10 MeV	>10 MeV	>100 MeV	>100 MeV
	Median	Target Quartile	Median	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.

- First broad assessment of SEP model skill
- Benchmark to quickly check if a new model preforms better than the SEPVAL median
- Provided in Chapter 8 of the technical paper

SEPVAL Post-eruptive Models				
Metrics	>10 MeV	>10 MeV	>100 MeV	>100 MeV
	Median	Target Quartile	Median	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).

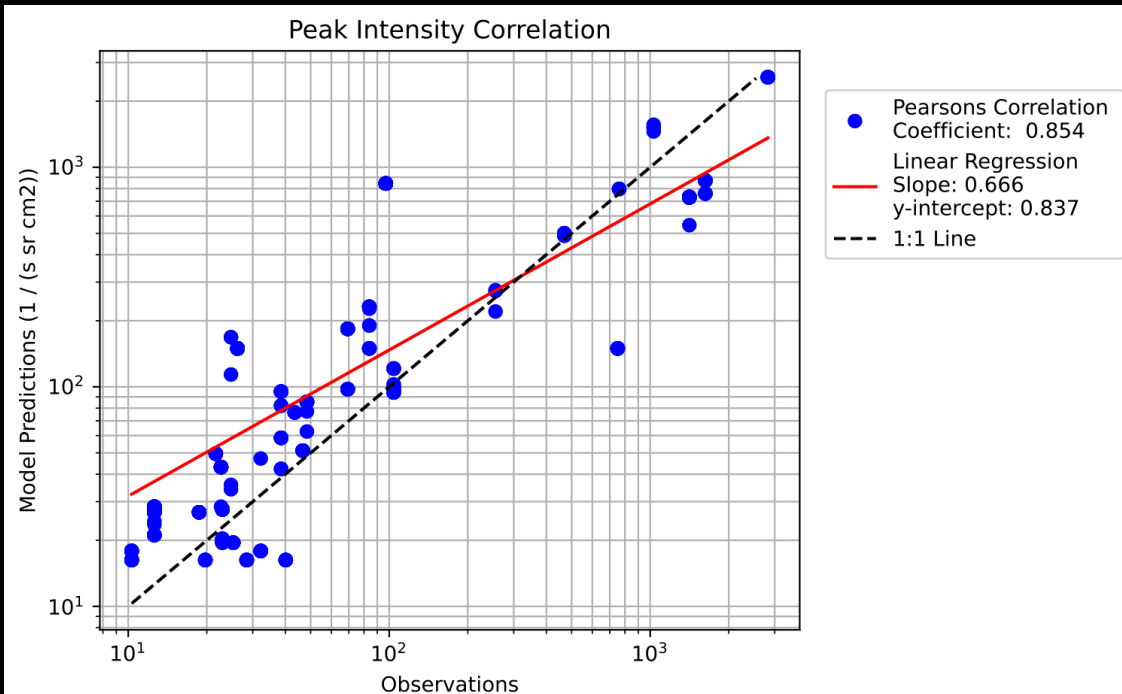
SEPVAL Takeaways

- SEP models have some skill and can generally discriminate between conditions that do and do not lead to SEP events
- “Best” model depends on the choice of metric
- Post-eruptive models have much higher skill than pre-eruptive models
- The median False Alarm Ratio for SEPVAL ranges from 0.03 – 0.46
 - 3% - 46% of “yes” forecasts are false alarms for a balanced dataset
 - This feels like reasonable performance if these rates were to translate to real-time forecasting...
- The poor correlation and large spread in peak flux predictions indicate that models do not reproduce event-to-event variability
- SEP models as a whole are lacking >100 MeV predictions and demonstrate lower skill at this energy range that is important to SRAG

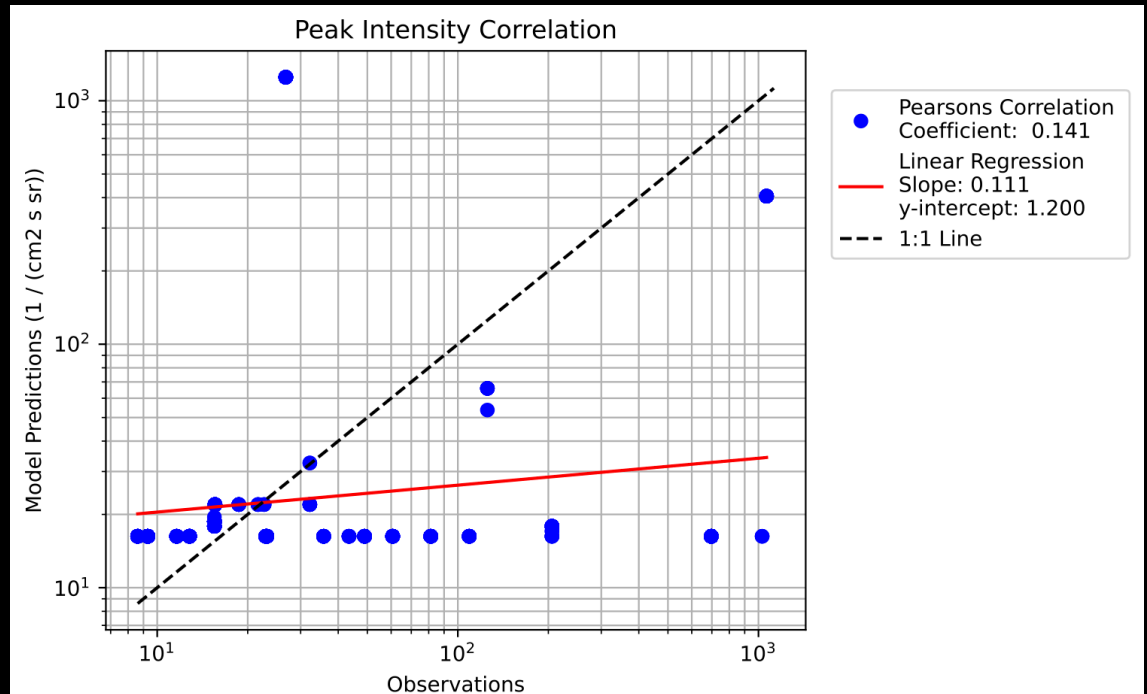
Are SEPVAL Results Representative of Real-time Forecasting Skill?

It is important to validate models in a real-time forecasting environment.

UMASEP-10 SEPVAL



UMASEP-10 SEP Scoreboard



SEP Scoreboards Timeline

• 2018

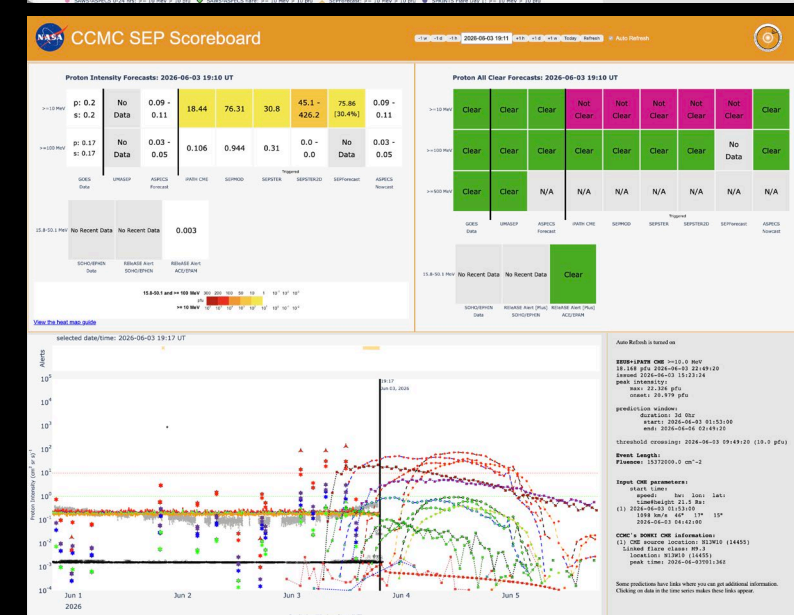
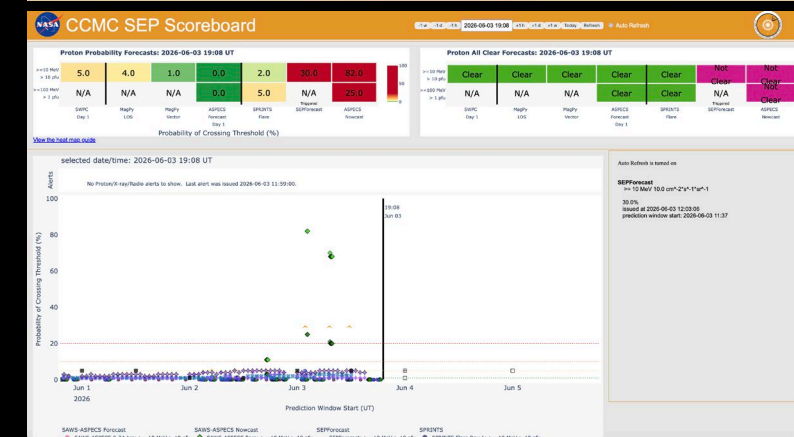
- Integrated Solar Energetic Particle Alert/Warning System (ISEP) collaboration established between SRAG and CCMC
- SEP Scoreboard planning and development begins
- "Mature" SEP models identified for first round of R2O2R development and onboarding to SEP Scoreboards

• 2020

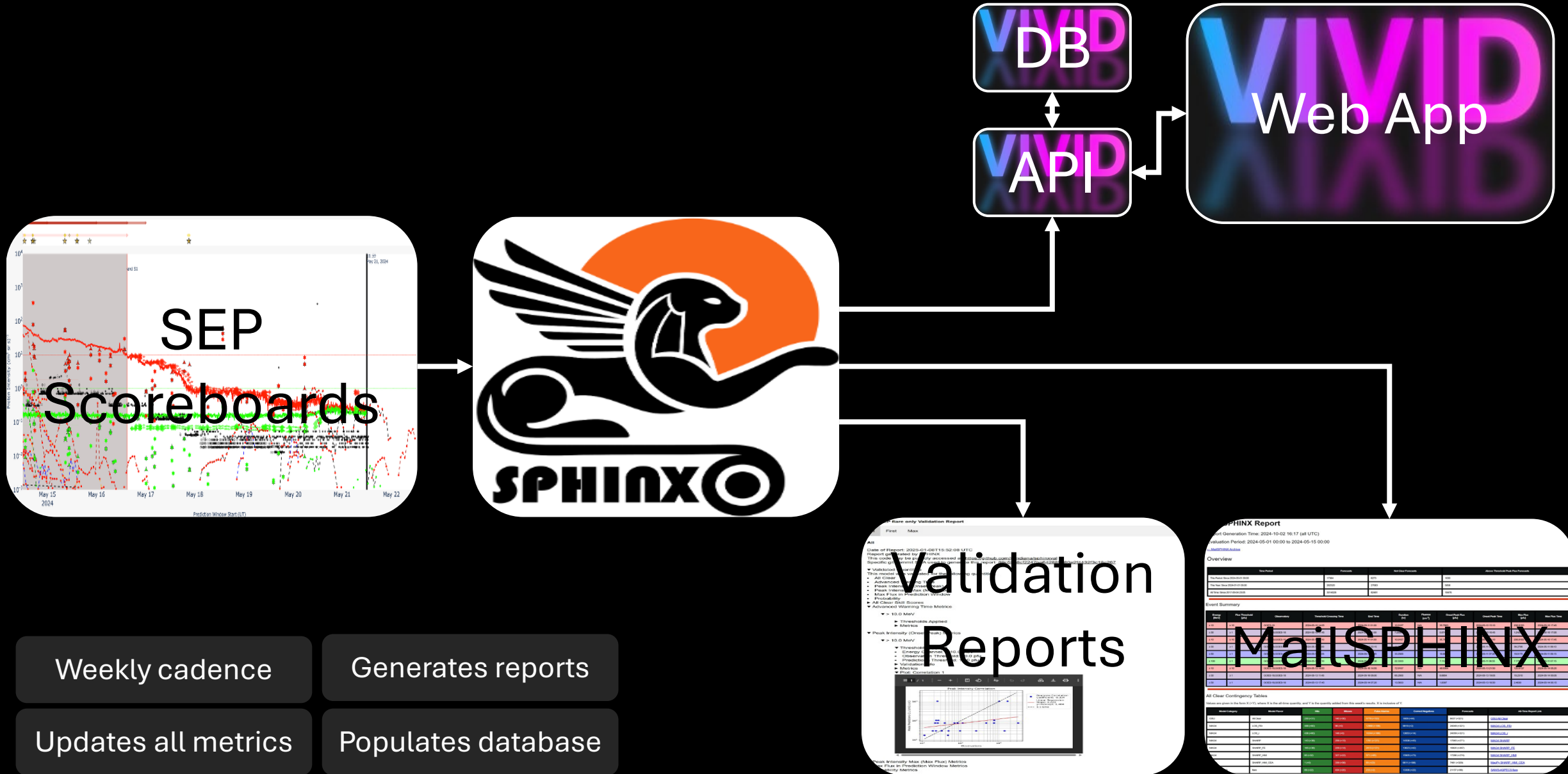
- M2M SWAO established, joins ISEP, provides human-in-the-loop support for the SEP Scoreboards
- March 2020 first models onboarded to the SEP Scoreboards and begin forecasting in real time

• 2026:

- NASA Technical Paper with **first validation results for 10 models**: Whitman, K., et al., *Validation of Solar Energetic Particle Forecasting Models for Space Radiation Operations with SPHINX and VIVID*. NASA/TP-20260000463 (2026)



Automated Validation Pipeline



Validation Reports

HTML **reports** generated for each model every time SPHINX runs

HTML **reports** generated and emailed monthly for the summary period, years-to-date, and all time

COMESAP flare only Validation Report

All First Max

All

Date of Report: 2025-01-06T15:52:08 UTC
Report generated by SPHINX.
This code may be publicly accessed at: <https://github.com/stindiana/sphinxval>
Specific git commit SHA used to generate this report: [89c5938c72247ba642889080e2f4432f3c18c267](https://github.com/stindiana/sphinxval/commit/89c5938c72247ba642889080e2f4432f3c18c267)

Validated Quantities

This model was validated for the following quantities.

- All Clear
- Advanced Warning Time
- Peak Intensity (Onset Peak)
- Peak Intensity Max (Max Flux)
- Max Flux in Prediction Window
- Probability
- All Clear Skill Scores
- ▼ Advanced Warning Time Metrics

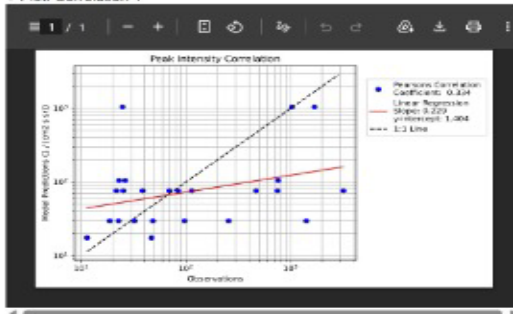
► > 10.0 MeV

- Thresholds Applied
- Metrics

Peak Intensity (Onset Peak) Metrics

► > 10.0 MeV

- ▼ Thresholds Applied
- Energy Channel: > 10.0 MeV
- Observation Threshold: 10.0 pfu
- Predictions Threshold: 10.0 pfu
- Validation Info
- Metrics
- ▼ Plot: Correlation 1



- Peak Intensity Max (Max Flux) Metrics
- Max Flux in Prediction Window Metrics
- Probability Metrics

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	Forecast	Not Clear Forecasts	Above Threshold Peak Plus Forecasts
This Period Since 2024-05-01 00:00	17064	6273	1930
This Year Since 2024-01-01 00:00	26520	27063	8959
All Time Since 2017-09-04 23:05	3014028	82481	19470

Event Summary

Energy [MeV]	Flux Threshold [pfu]	Observatory	Threshold Crossing Time	End Time	Duration [s]	Fluence [mfu*s]	Onset Peak Flux [pfu]	Onset Peak Time	Max Flux [pfu]	Max Flux Time
3-10	2-10	GOES-18	2024-05-10 14:05	2024-05-11 01:00	10,9167	N/A	35.7803	2024-05-10 15:10	206.9193	2024-05-10 17:45
3-30	2-1	GOES-18, GOES-18	2024-05-10 17:45	2024-05-11 01:00	7,2000	N/A	0.6740	2024-05-10 19:45	1.2403	2024-05-10 17:55
3-10	2-10	GOES-18, GOES-18	2024-05-11 02:10	2024-05-11 01:00	10,9167	N/A	35.7803	2024-05-10 15:10	206.9193	2024-05-10 17:45
3-30	2-1	GOES-18, GOES-18	2024-05-11 02:05	2024-05-12 22:15	44,2500	N/A	27.2808	2024-05-11 07:26	36.0796	2024-05-11 09:10
3-30	2-1	GOES-18, GOES-18	2024-05-11 02:05	2024-05-12 11:20	33,2500	N/A	16.8888	2024-05-11 07:26	19.9175	2024-05-11 09:15
3-100	2-1	GOES-18, GOES-18	2024-05-11 02:10	2024-05-12 00:00	22,3333	N/A	7.1125	2024-05-11 06:55	7.7752	2024-05-11 07:15
3-10	2-10	GOES-18, GOES-18	2024-05-13 14:00	2024-05-16 14:05	72,9167	N/A	48.8394	2024-05-13 21:30	158.9937	2024-05-14 00:20
3-30	2-1	GOES-18, GOES-18	2024-05-13 14:45	2024-05-16 06:00	45,2500	N/A	8.8894	2024-05-13 19:05	15.2316	2024-05-14 00:05
3-30	2-1	GOES-18, GOES-18	2024-05-13 17:45	2024-05-14 07:20	13,9603	N/A	1.8397	2024-05-13 18:50	2.4000	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)	142(+83)	3576(+103)	1839(+46)	9437(+321)	GSU All Clear
MAGA	LOS_FE	488(+40)	85(+8)	1080(+186)	8919(+43)	3494(+321)	MAGA LOS_FE
MAGA	LOS_J	106(+40)	148(+8)	1034(+186)	1283(+14)	3499(+321)	MAGA LOS_J
MAGA	SWARP	143(+36)	228(+15)	1781(+121)	1400(+40)	1706(+271)	MAGA SWARP
MAGA	SWARP_FE	165(+36)	228(+15)	2413(+121)	1363(+40)	1800(+307)	MAGA SWARP_FE
MAGA	SWARP_HH	85(+30)	37(+20)	571(+46)	1585(+70)	1726(+274)	MAGA SWARP_HH
MegaY	SWARP_HH_CSA	1(+0)	38(+46)	8(+20)	671(+186)	788(+326)	MegaY SWARP_HH_CSA
SAWSASPECTS	flr	89(+20)	64(+20)	201(+5)	1006(+20)	2117(+46)	SAWSASPECTS flr

VIVID

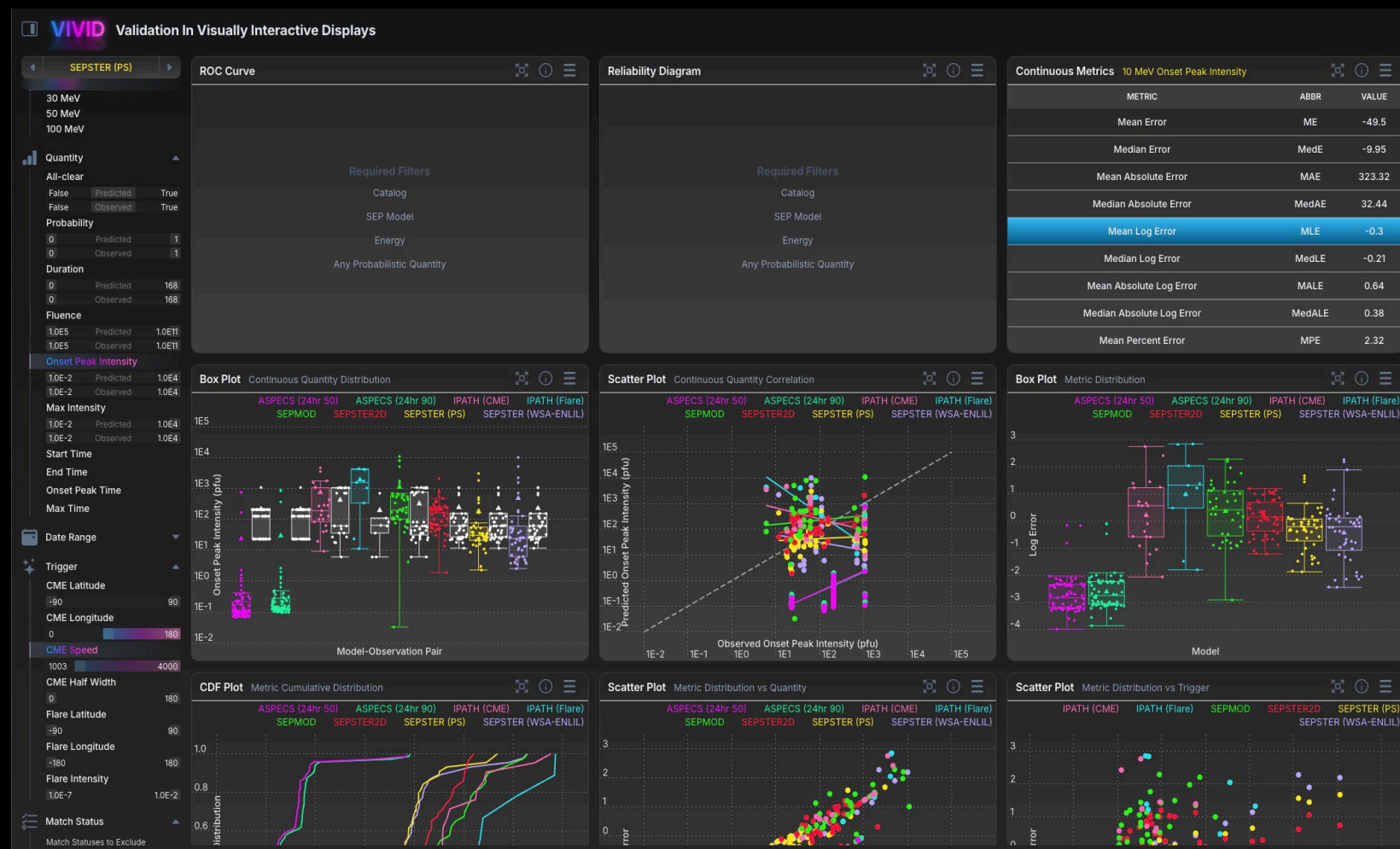
<https://ccmc.gsfc.nasa.gov/vivid/>

Displays SEP Scoreboard
validation results produced
by SPHINX

Users can **filter** by any
quantity or trigger

Allows **cross-model**
comparison for the same set
of SEP events

Metrics are **recalculated**
based on the filtered subset
of predictions and events



SEP Scoreboards Dataset and Models

Climatological, Imbalanced Dataset

- March 2020 to December 2024
 - 37 SEP events, >10 MeV, 10 pfu
 - 9 energetic SEP events, >100 MeV, 1 pfu
- True climatology, highly imbalanced
- 11 models
- SWPC as a model:
 - 1 Day (pre-eruptive)
 - Warnings (post-eruptive)

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).

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

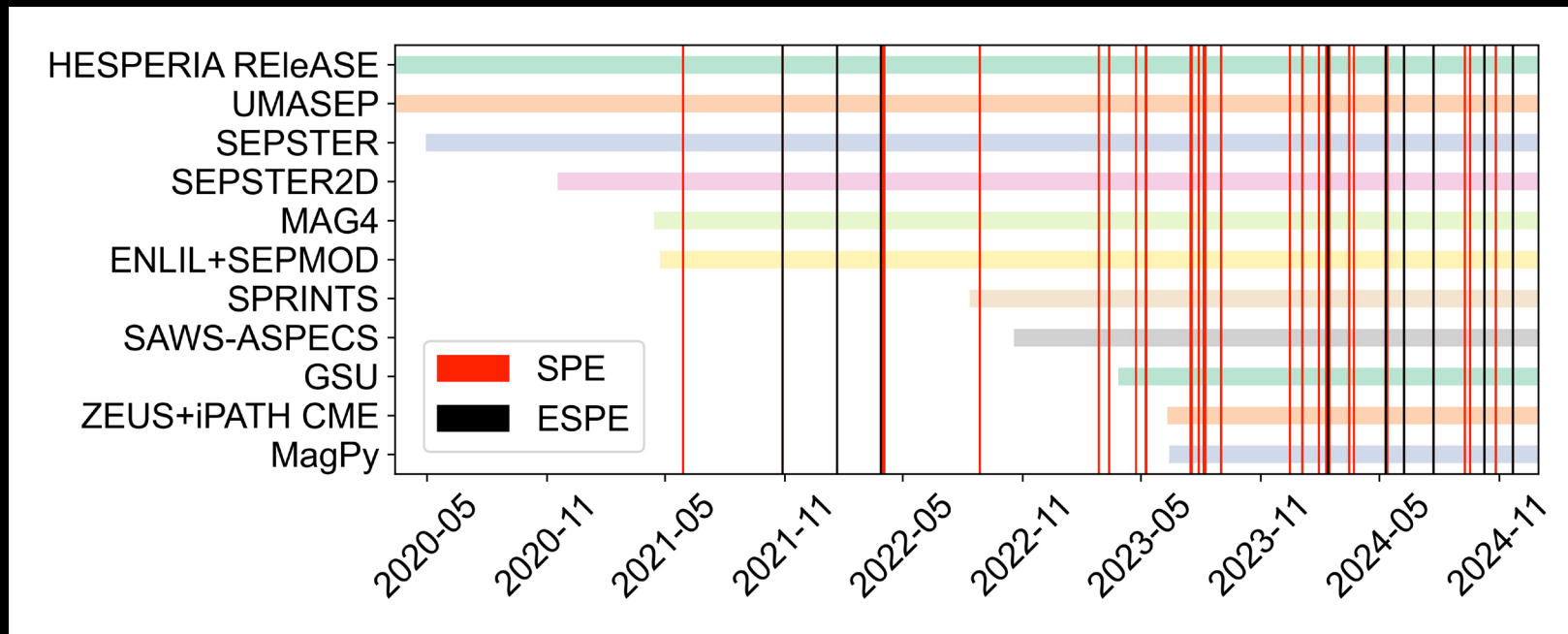
Philip Quinn
Leidos, Houston, TX

Luke Stegeman
Leidos, Houston, TX

<https://ntrs.nasa.gov/citations/20260000463>

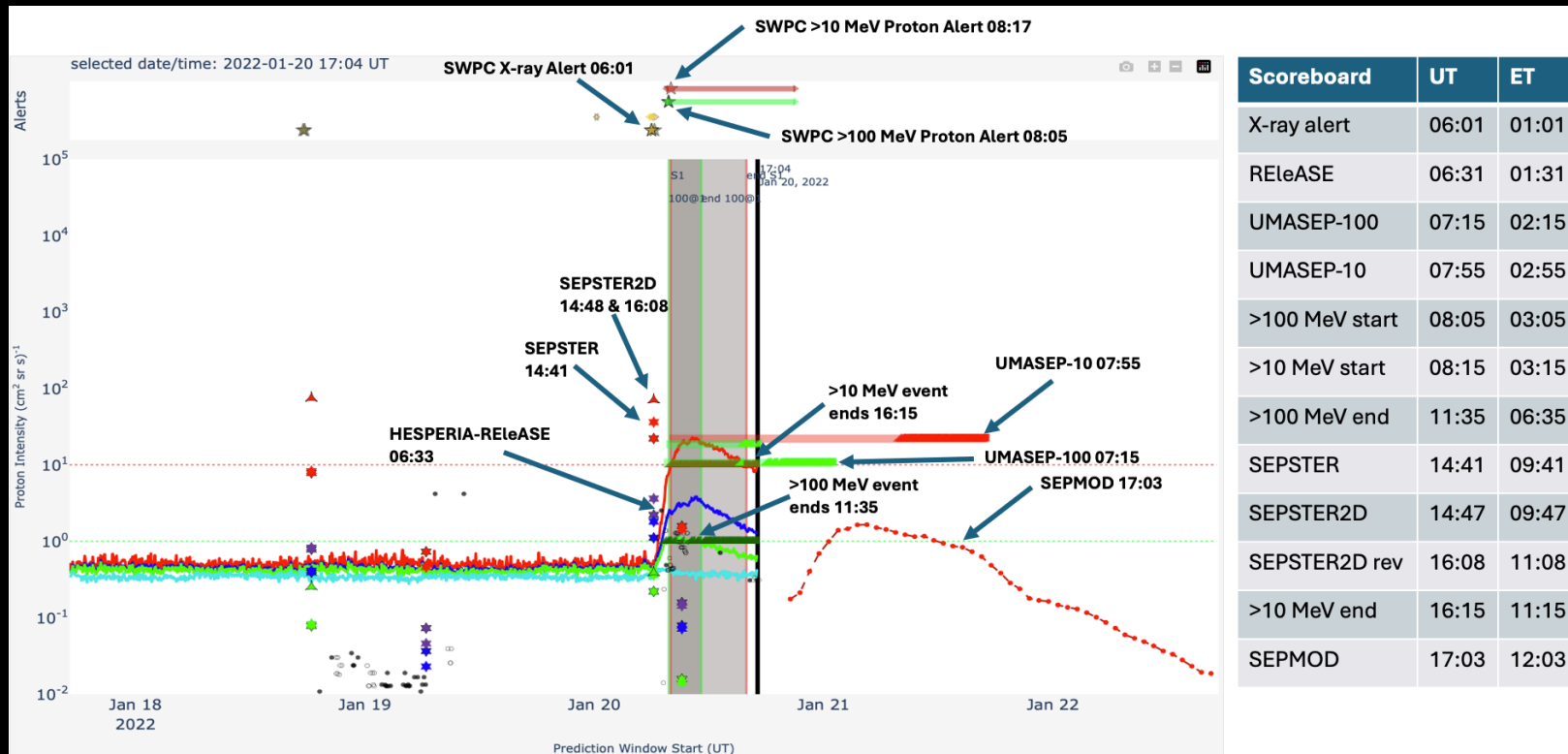
SEP Scoreboards Model Coverage

- Each model has a different evaluation time period and climatology
- Each model has a different cadence: pre-eruptive and post-eruptive
- Millions of forecasts have been submitted to the SEP Scoreboards (~150,000/month)
 - SPHINX manages forecast revisions, duplicates, and corrupt or incomplete forecasts



Validation of the Real-time System

- SEP Scoreboard validation metrics reflect:
 - Availability and quality of real-time measurements (data gaps, data latency)
 - Underlying model approach
 - SEP Scoreboard system as a whole



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

SEPSTER Results

SEP Scoreboard vs SEPVAL

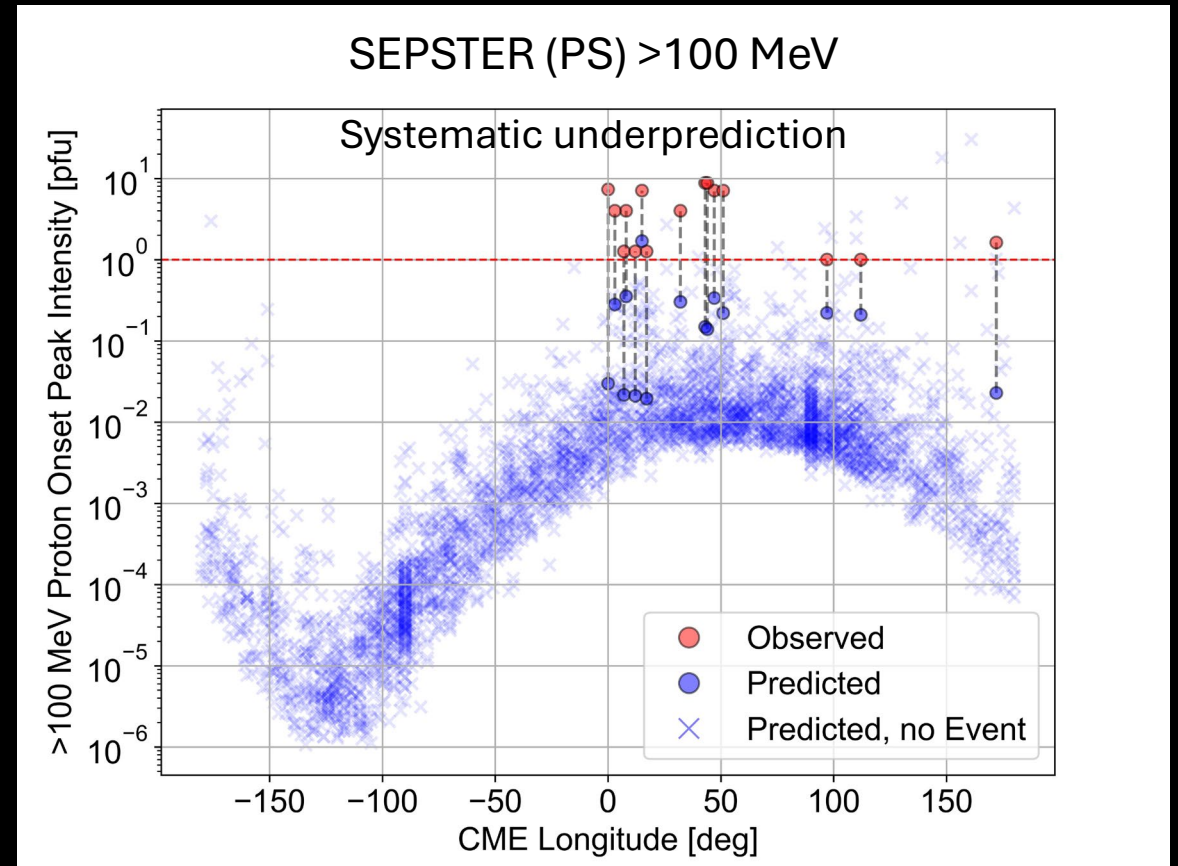
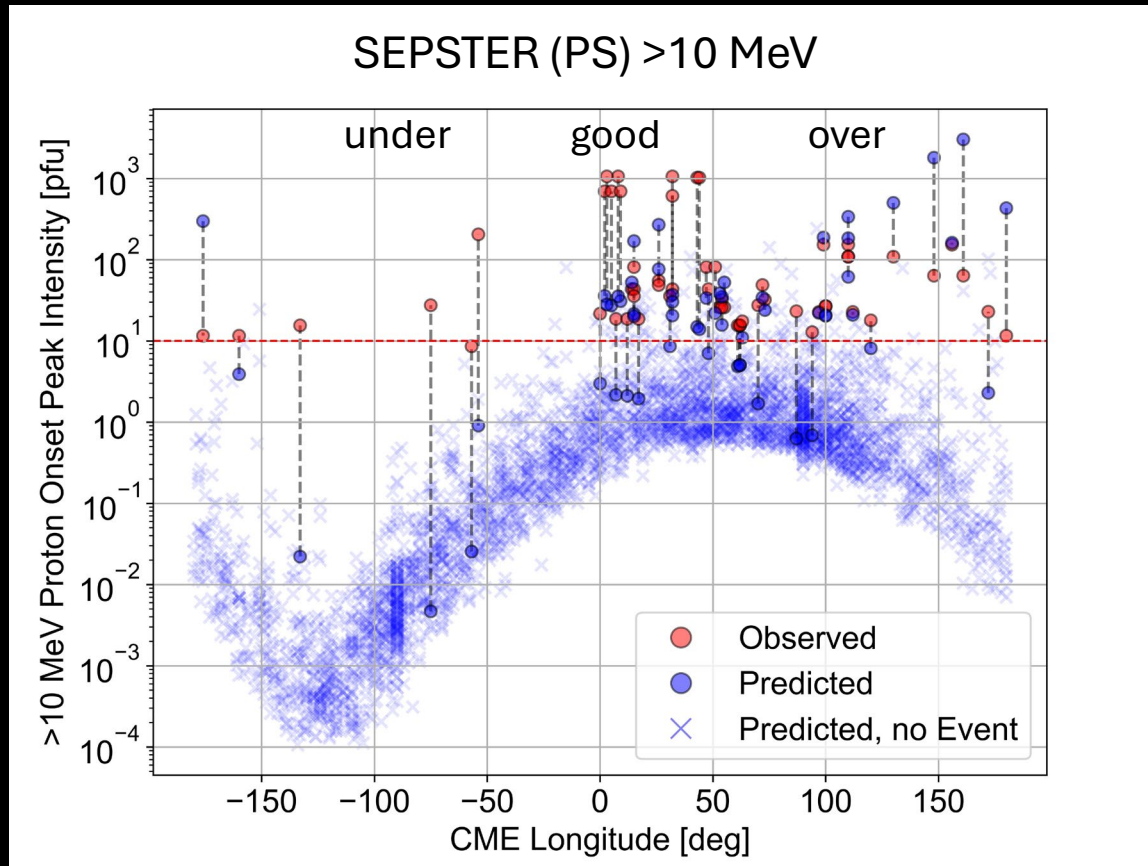
SEPSTER All Clear	SEPVAL		Scoreboard	
	> 10 MeV (<i>N</i> = 61)	> 100 MeV (<i>N</i> = 62)	> 10 MeV (<i>N</i> = 3763)	> 100 MeV (<i>N</i> = 3067)
Percent Correct	0.74	0.82	0.98	0.992
Hit Rate	0.65	0.44	0.62	0.08
False Alarm Rate	0.17	0.043	0.018	0.0043
False Alarm Ratio	0.20	0.22	0.64	0.93
Bias	0.81	0.44	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.

- Numerical values of skill Scores for SEPVAL (balanced) and SEP Scoreboards (highly imbalanced) cannot be compared
- Some scores are overwhelmed by correct negatives (Percent Correct, False Alarm Rate, TSS)
- Some scores are more sensitive to hits and false alarms (Hit Rate, False Alarm Ratio, Threat Score, HSS)
- **SRAG needs false alarms to be minimized while retaining hits, so important to use scores sensitive to false alarms**
- Percent Correct is very high due to correct negatives, $(h+c)/N$
- Hit Rate remains consistent for >10 MeV, $h/(h+m)$
- Hit Rate significantly lower for >100 MeV
- False Alarm Rate goes down significantly, $f/(f+c)$
- False Alarm Ratio increases significantly, $f/(f+h)$

SEPSTER Results

Systematic Performance



SEP Scoreboard Validation Caveat

- **SEP Scoreboard metrics will change compared to the NASA Technical Report**
- SPHINX has been upgraded to provide **error bars** on the metrics
- SPHINX has been updated to use **known SEP, CME, and flare associations**
 - Upgrades to FetchSEP to generate the CLEAR SEP Dataset enabled this capability in SPHINX

<https://ccmc.gsfc.nasa.gov/swxcoe/clear/benchmark.php>

**Community Coordinated Modeling Center**

Search

AboutModelsSimulation ServicesValidationCommunity SupportSpace WeatherTools

[← Back to CLEAR \(top level\)](#)



CLEAR SEP Benchmark Dataset v2.0

Core SEP event lists, supporting products, and documentation for the CLEAR benchmark dataset. SEP events identified above operational thresholds for >10, >100, >30, and >50 MeV and above GOES detector background. The lists contain values from the primary GOES spacecraft at the time of the SEP event.

Core Products

Operational List

Integral GOES fluxes as-is from NOAA archives; primary list for model development and validation against historical operational values (GOES-06 to GOES-19, Jan 1986–Sep 2025).

[GOES_integral_PRIMARY.1986-02-03.2025-09-10_sep_events.csv](#) [1.94 MB]

Energy-bin Calibrated List

Differential, background-subtracted, cross-calibrated GOES fluxes for science and dosimetry studies; integral fluxes estimated from calibrated channels (GOES-13 and GOES-15, 2010–2020).

[GOES_differential_energy_bin_calibrated_PRIMARY.2010-08-02.2017-09-20_sep_events.csv](#) [502.58 KB]

Code & Provenance

The CLEAR benchmark lists and supporting products are generated using the fetchsep tools (idsep/opsep) for identifying SEP elevations and analyzing individual events.

[fetchsep code repository \(GitHub\)](#)

[CLEAR Benchmark v2.0 tagged release](#)

For now this page highlights the two primary CLEAR SEP lists and the full dataset. Future updates may add direct links between individual rows and the key plots generated for each event.

Event Definitions	Number of Events	With Associations
>10.0 MeV 10.0 pfu	263	263
>10.0 MeV above background	532	484
>100.0 MeV 1.0 pfu	88	88
>100.0 MeV above background	162	159
>30.0 MeV 1.0 pfu	257	252
>30.0 MeV above background	353	338
>50.0 MeV 1.0 pfu	160	158
>50.0 MeV above background	290	279

Table 1: Number of SEP events per event definition in the Operational SEP List in the CLEAR SEP Benchmark Dataset Version 2.0.

SEP Scoreboard Takeaways

- Best All Clear
 - UMASEP-10: Hit Rate = 69%, False Alarm Ratio = 63%, HSS = 0.46
 - SEPSTER: Hit Rate = 62%, False Alarm Ratio = 64%, HSS = 0.44
 - HESPERIA REleASE SOHO-60 min (15.8–39.8 MeV) to GOES >10 MeV, 10 pfu: Hit Rate = 88%, False Alarm Ratio = 42%, and HSS = 0.70.
- Model skill and availability (AWT) not yet sufficient for operational use
- State-of-the-art performance for real-time forecasting is provided in the NASA TP and attached as an Appendix to this presentation

SRAG proposes a new skill score called **False Alarm Event Ratio (FAER)**, pronounced “fear”, defined as:

$$\text{FAER} = \text{false alarms} / (\text{hits} + \text{misses})$$

This ratio represents the number of false alarms compared to the number of observed SEP events and has a simple intuitive utility as models with high FAER are impossible to trust. A goal of $\text{FAER} \leq 1$ provides a constraint on the False Alarm Rate and Ratio.

UMASEP FAER = 1.17 (3.84 raw), SEPSTER FAER = 1.11, REleASE FAER = 0.62

Back Ups

Commonly Confused or Abused Words

- **Validation versus Verification** - used in opposite senses in different communities
 - Is the software or system is working as expected?
 - Does the output represent reality (observations)?
 - I use the term “validation” to mean evaluating a model’s performance by comparing predictions to observations
 - For the SEP Scoreboards, which collect forecasts in real time, the term “forecast verification” is used in the meteorological community (I will still say validation)
- **Operational**
 - Operations requires rigorous methodologies, workflows, and systems which depend on safety margins, fallbacks, and redundancy and is supported 24/7
 - A model is operational if it achieves RL 9 - fully validated, tested, demonstrated to be robust, and running in an operational environment using operational inputs (or inputs that have been deemed by the user to have an acceptable level of reliability)
- **Real time**
 - A model runs in real time if it is a fully automated system that uses inputs that are available in real time

SEPVAL Timeline

- **SHINE 2018**
 - CCMC and SWPC SEP validation challenge at SHINE
- **SHINE 2019 and ISWAT 2020**
 - SHINE challenge expanded to 9 SEP events
 - First prototype of FetchSEP to prepare observed SEP values
- **2021**
 - Begin development on SEP validation prototype
- **SHINE 2022 and ISWAT 2022**
 - Add 14 non-event periods to test for false alarms
<https://helioshine.org/sep-model-validation-challenge/>
<https://ccmc.gsfc.nasa.gov/assessment/topics/SEP/campaign2020.php>
- **SEPVAL 2023, mini-ISWAT, ESWW 2023**
 - Extend challenge to 33 SEP events and 30 non-events (SEPVAL Dataset)
<https://zenodo.org/records/15555244>
 - Begin development of SPHINX <https://github.com/ktindiana/sphinxval>
 - Begin development of VIVID <https://ccmc.gsfc.nasa.gov/vivid/>
- **2026**
 - Whitman et al. (2026). *Validation of Solar Energetic Particle Forecasting Models for Space Radiation Operations with SPHINX and VIVID*. NASA NTRS.
<https://ntrs.nasa.gov/citations/20260000463>



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*

SEP Scoreboard Maximum Flux Forecasts

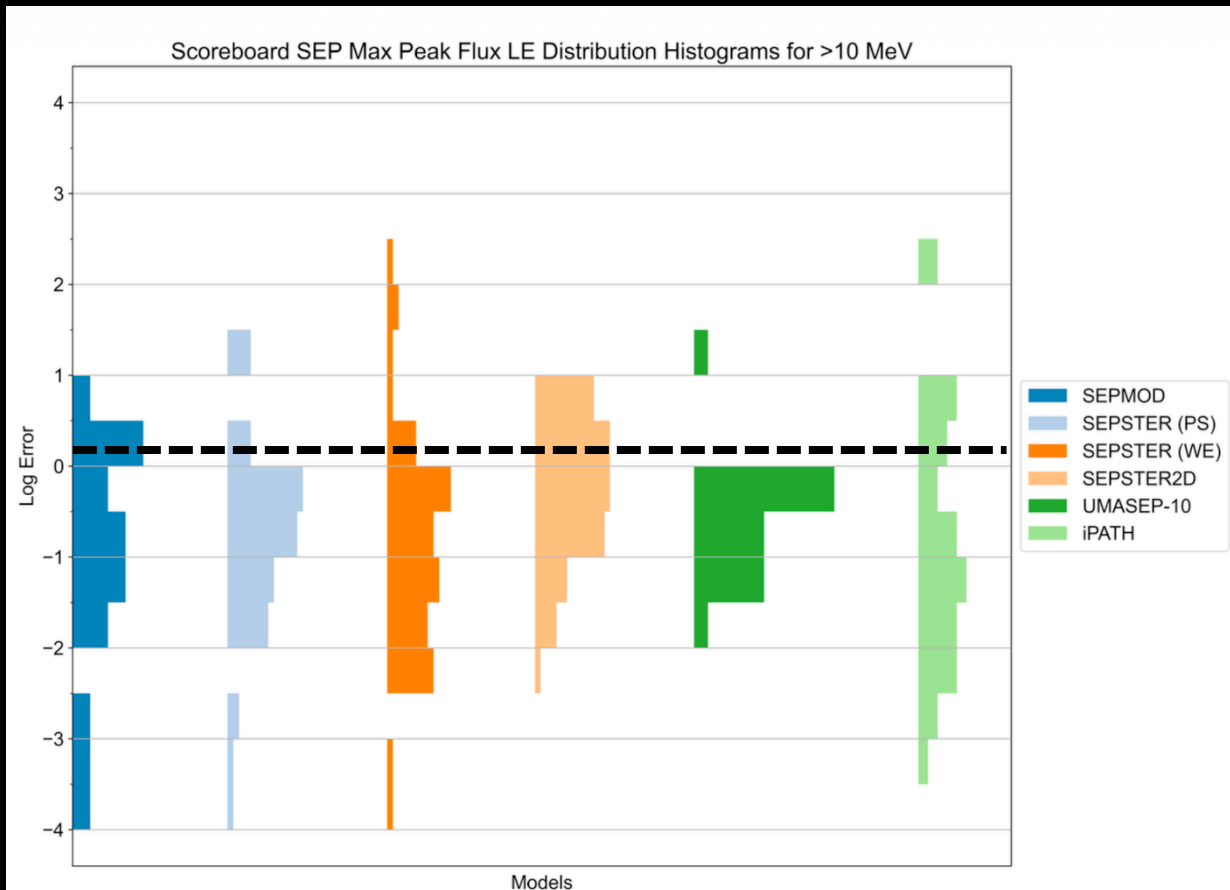
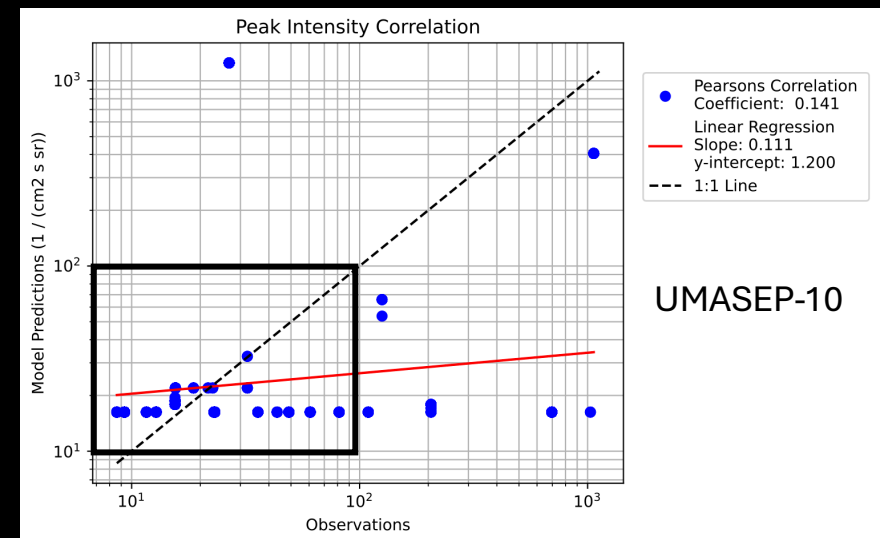


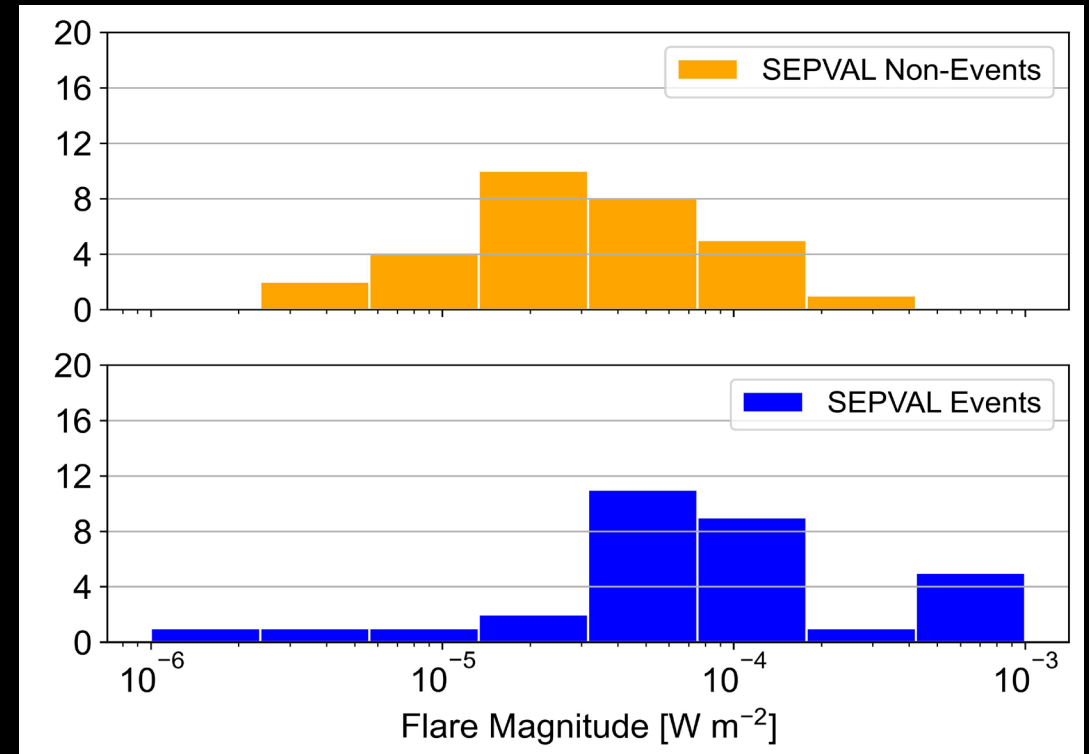
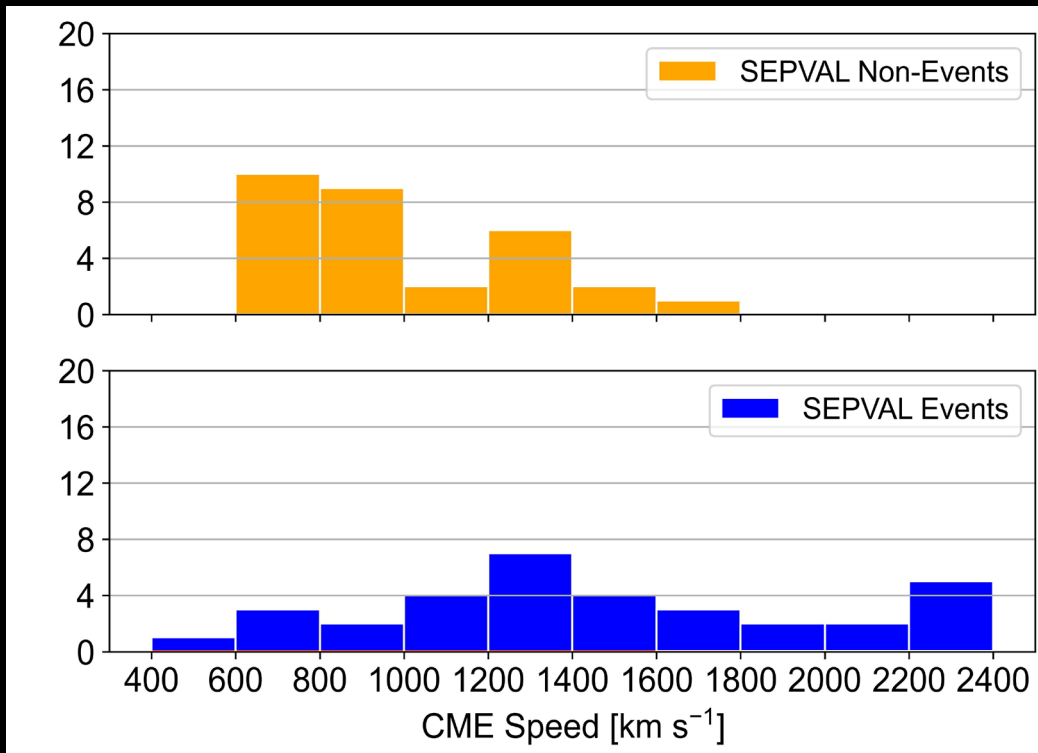
Figure 5.40: Summary histograms of SEP Scoreboard Max Flux log error for forecasts associated with >10 MeV integral proton flux with 10 pfu flux threshold.

- Error in forecasts up to 2 orders of magnitude
- Models do not capture event-to-event variability
- Too unreliable to use for decision-making



SEPVAL Event Set

- Both SEP event and non-event time periods were associated with strong flares and fast CMEs
- Can models discriminate between conditions that lead to SEP events and those that don't?



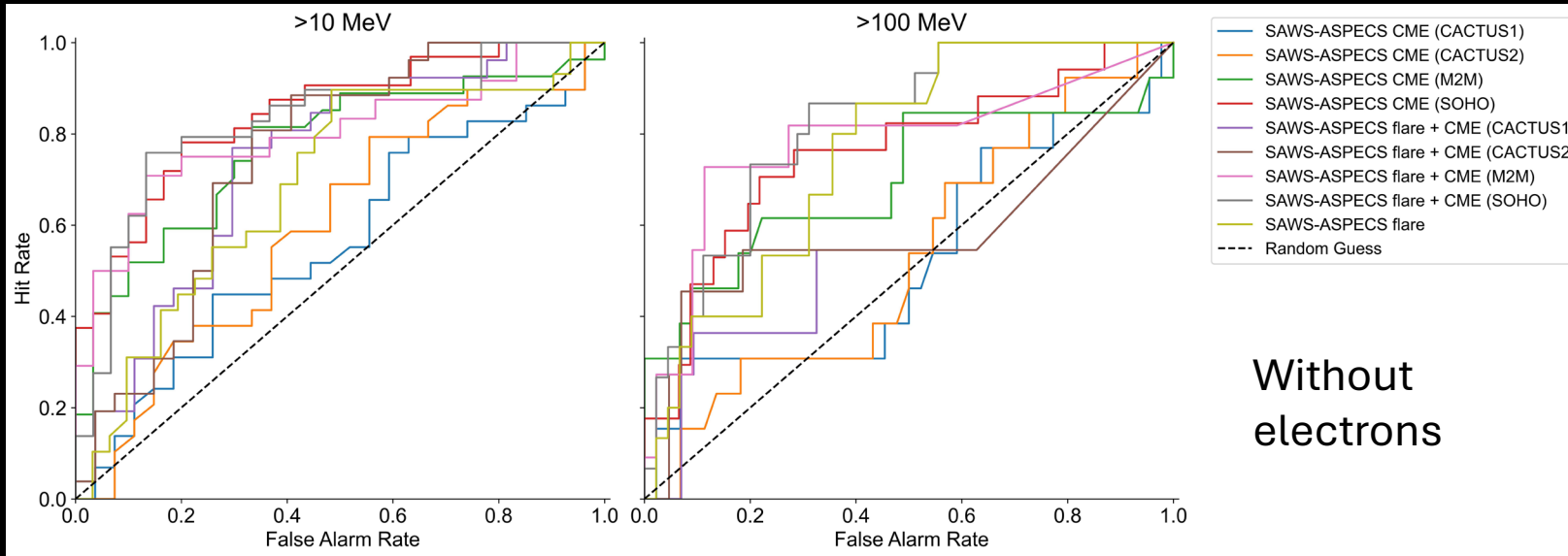
SEPVAL Maximum Flux Results

>10 MeV, 10 pfu for Post-Eruptive Models

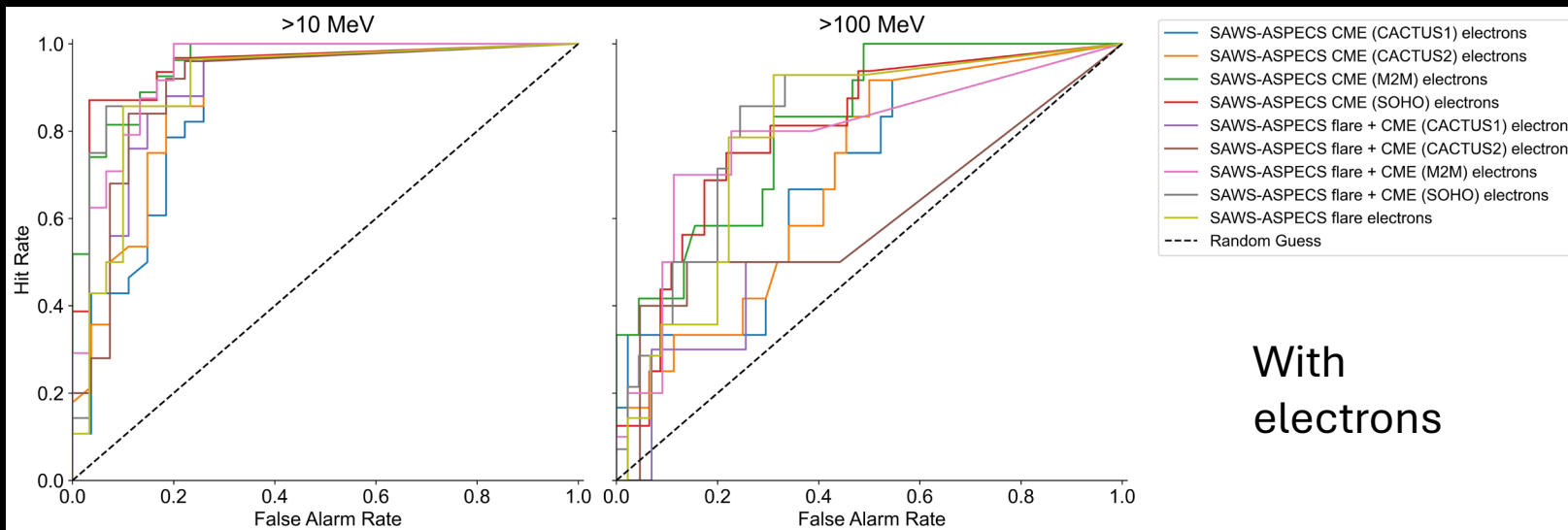
Model	Pearson Correlation Coefficient (Log)	Spearman Correlation Coefficient (Linear)	Median Log Error (MedLE)	Mean Absolute Log Error (MALE)	Ranked by Median Absolute Log Error (MedALE)	Root Mean Square Log Error (RMSLE)
UMASEP-10	0.89	0.88	0.01	0.28	0.26	0.36
COMESep flare only	0.42	0.42	-0.13	0.54	0.27	0.77
ADEPT-AFRL 1hr	0.19	0.33	-0.06	0.56	0.30	0.90
COMESep flare+CME	0.40	0.34	-0.17	0.57	0.43	0.74
SEPSTER2D CME	0.58	0.59	-0.04	0.59	0.50	0.77
SEPSAT	0.34	0.33	-0.09	0.70	0.56	0.89
SFS-Update	0.41	0.42	-0.38	0.66	0.57	0.90
ZEUS+iPATH_CME	0.07	0.24	-0.22	1.34	0.77	2.17
ENLIL+SEPMOD	0.29	0.32	-0.60	1.23	0.79	1.71
SEPSTER (Parker Spiral)	0.55	0.51	-0.71	0.81	0.81	1.01
SAWS-ASPECS flare + CME (SOHO) electrons 50%	0.27	0.49	0.58	0.94	0.84	1.16
SAWS-ASPECS flare + CME (SOHO) 50%	0.29	0.48	0.55	0.93	0.88	1.13
SAWS-ASPECS flare 50%	0.39	0.40	-0.67	1.10	1.21	1.39
SAWS-ASPECS flare electrons 50%	0.38	0.36	-0.81	1.16	1.21	1.44
SAWS-ASPECS flare electrons 90%	0.42	0.37	0.66	1.46	1.59	1.67
SAWS-ASPECS flare 90%	0.43	0.40	0.96	1.51	1.62	1.69
SAWS-ASPECS flare + CME (SOHO) 90%	0.27	0.48	1.90	1.82	2.05	1.98
SAWS-ASPECS flare + CME (SOHO) electrons 90%	0.25	0.49	1.93	1.80	2.05	1.96
SAWS-ASPECS CME (SOHO) electrons 90%	0.36	0.21	-2.10	2.21	2.10	2.34
SAWS-ASPECS CME (SOHO) 90%	0.44	0.23	-2.12	2.28	2.12	2.39
SAWS-ASPECS CME (SOHO) electrons 50%	0.38	0.21	-2.16	2.30	2.16	2.41
SAWS-ASPECS CME (SOHO) 50%	0.43	0.21	-2.17	2.34	2.17	2.45

How have the SEPVAL results
been used?

SAWS-ASPECS Systematic Study with SEPVAL

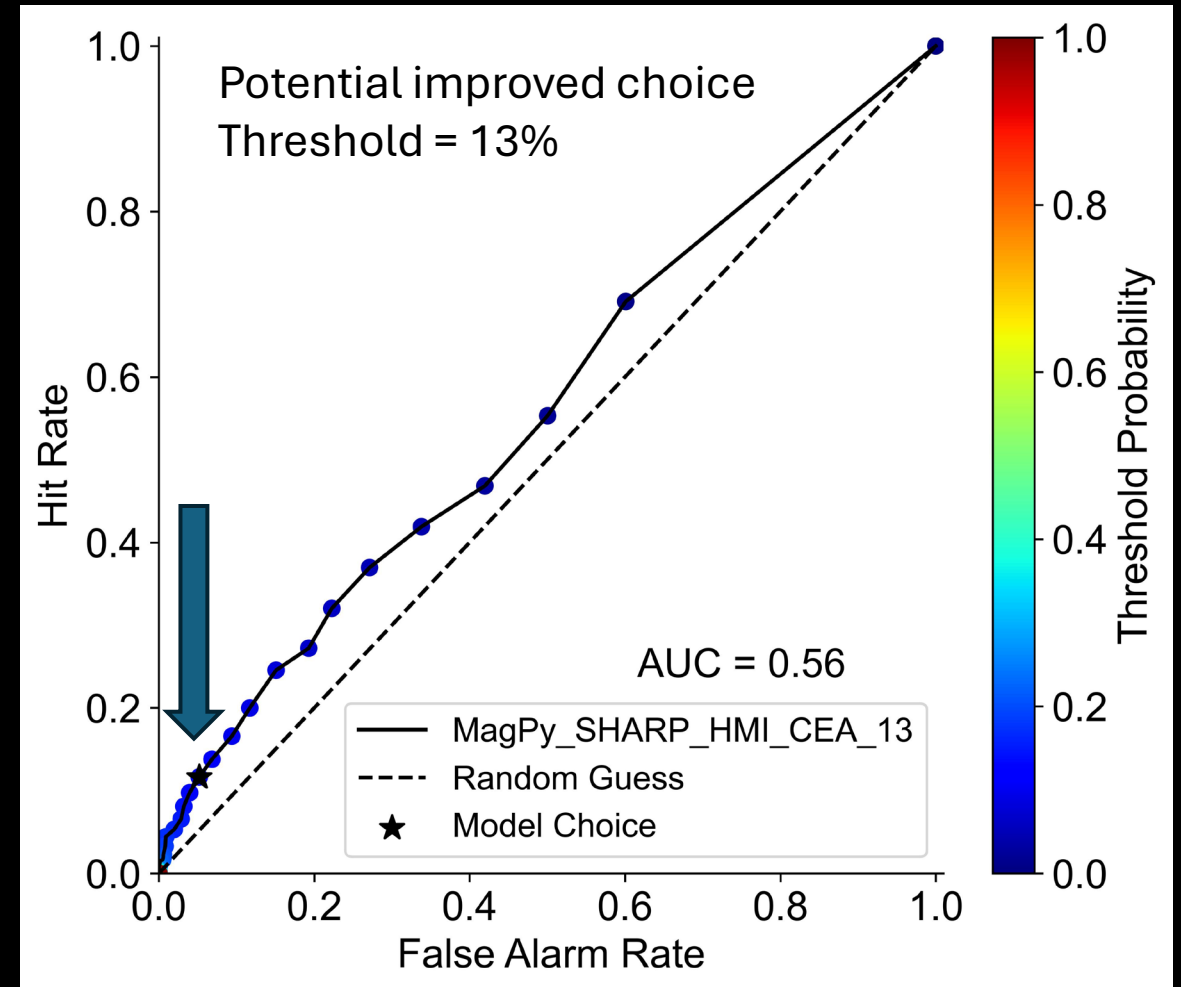
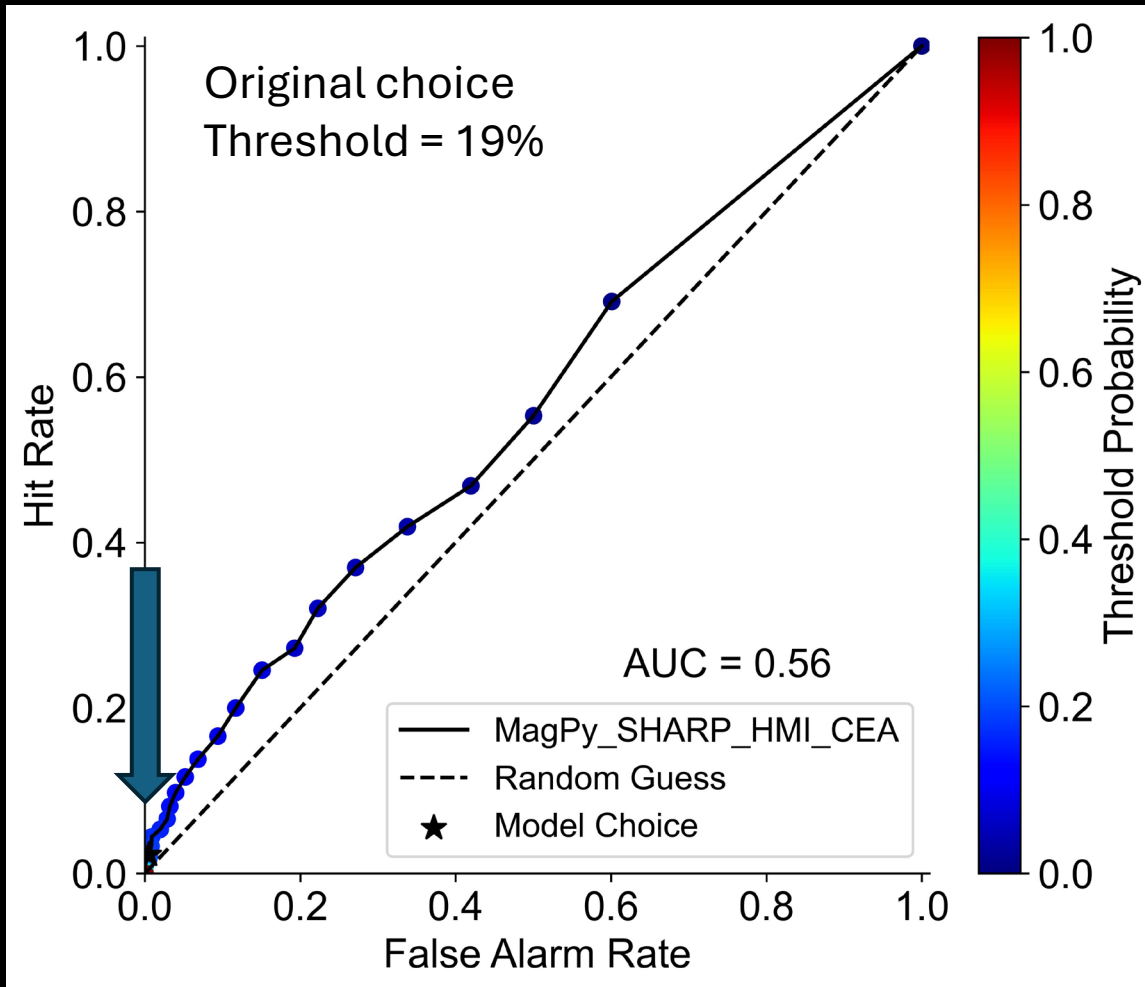


- Multiple versions of their model using different combinations of inputs
- Used all CME parameters: CDAW, DONKI, CACTus
- With and without electrons



MagPy HMI SHARP All Clear Performance

Different choice of probability threshold will improve MagPy's All Clear performance



Identify Gaps

SEPVAL Median All Clear Scores	SEPVAL post-eruptive		SEPVAL pre-eruptive	
	>10 MeV	>100 MeV	>10 MeV	>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.

- Lacking pre-eruptive models that predict >100 MeV
- Skill for >100 MeV is lagging behind >10 MeV

SEPVAL Median Flux Metrics	Onset Peak		Maximum Flux	
	>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.

- Median performance of peak flux prediction is ~ an order of magnitude
- Correlation between predicted peak and observed peak is very low
- All scores are worse for >100 MeV