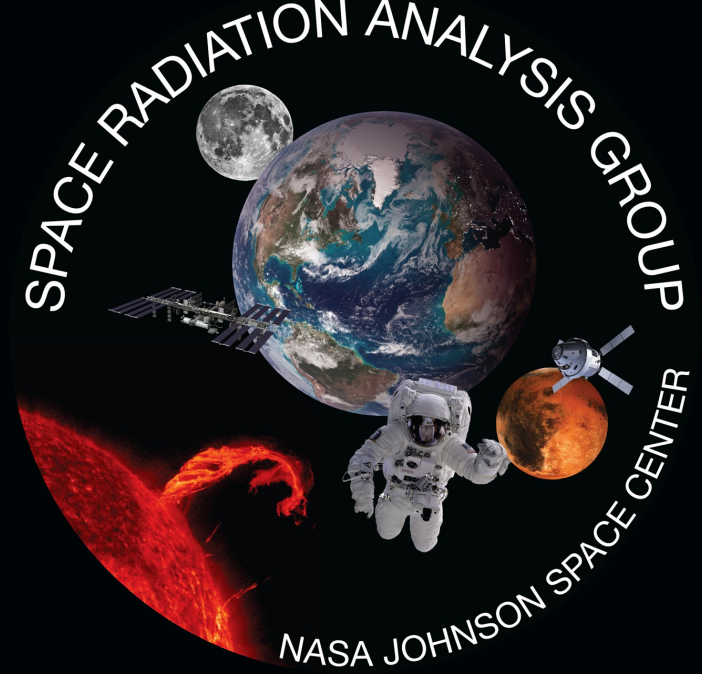
SEPVAL 2023 Post Analysis Results

Kathryn Whitman, Philip Quinn, Ricky Egeland, Luke Stegeman, Clayton Allison

NASA JSC SRAG

With contributions from many research and operational institutions

April 2024



And the 17+ participating model developers' institutions

Overview

- This presentation is intended to inform all those who attended the SEPVAL 2023 working meetings of a post analysis that has been performed with the submitted forecasts
  - Updated version of SPHINX with bug fixes, added metrics, and improvements
  - SPHINX code tagged on github with SEPVAL_2023_Post_Analysis: https://github.com/ktindiana/sphinxval
- A summary of the SEPVAL efforts and challenge event periods is described
- The full set of forecasts that were submitted is listed
- Anonymized examples of the metrics are shown here to give an idea of the spread in performance
- Each model developer will receive a separate package with their model's results along with summary plots showing their model's performance in comparison with the other participants (anonymized)
- We will eventually deploy a publicly accessible version of VIVID loaded with the SEPVAL results so that users may explore model performance interactively

Overview of the ISEP Validation Effort

The ISEP validation effort began in 2018/2019

Prior to that, many developers of Solar Energetic Particle (SEP) models did not focus on validation and did not have a quantitative sense of how their models performed

NOAA SWPC is SRAG's operational space weather forecasting service and their human-in-the-loop performance was largely unquantified until Bain et al. 2021

Thus, at the start of ISEP, the state-of-the-art of SEP forecasting performance was unclear because it had not been assessed

Through ISEP efforts and a series of validation challenges in partnership with the research community, the SPHINX validation framework was developed

The latest, most comprehensive SEPVAL challenge, along with the maturation of the SPHINX tool, has now provided a quantitative understanding of the state-of-the-art of SEP model performance

SEPVAL 2023 US and Europe Working Meetings

- Following a multi-year validation effort through the SHINE, ISWAT, and ESWW workshops, we established the SEPVAL workshops which were focused on validation and SEP forecasting in ops

- Assess SEP model performance
- Establish standards
- Develop a generalized framework for model validation
- Operational use of SEP models
- NASA, NOAA, ESA infrastructure being developed for the R2O transitioning of models to ops

Broader Goals of this Continuing Validation Effort

Generate an assessment of the **state-of-the-art** of the SEP modeling field.

Facilitate a culture shift in the scientific community to put more emphasis on thorough, quantitative validation

Develop best practices and standards and **communicate the forecasted quantities** most useful from end-user (SRAG, etc) and space weather provider (SWPC, ESA, etc) perspectives.

Simulated or actual real time validation to provide operators with a quantitative understanding of model performance by assessing the model **in the real-time environment in which it is used**. This may be different from the validation that is reported in the literature.

Sharing results of the challenge efforts and the SEP Scoreboards with model developers to motivate improvements (R2O cycle)

Organization of the SEPVAL Challenge

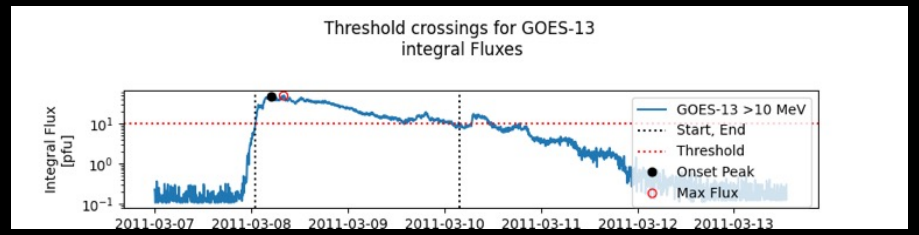
- **SEPVAL organizers:**
  - Provide a list of challenge time periods and triggers (flares, CMEs)
    - *M2M ensured quality CME measurements by checking all 3D CME parameters in DONKI provided for this challenge while also providing additional information*
  - Define rules of participation to encourage modelers to produce forecasts in a real time-like scenario
  - Perform the validation using SPHINX
  - Make the validation results available to attendees (R2O2R)

- **SEP model developers:**
  - Provide forecasts and supplementary information
  - Follow the rules of participation
  - Provide feedback about the forecast/prediction process and the validation results

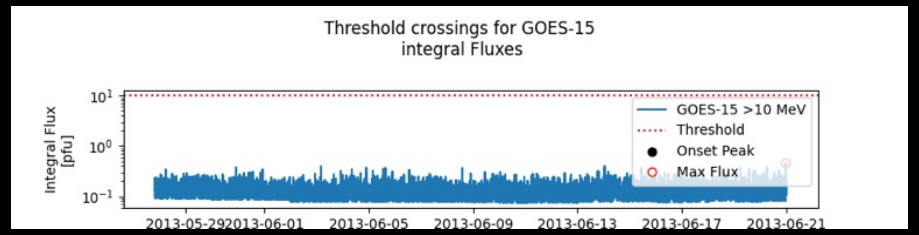
SEP Challenge Events Overview

- **33 SEP events** between 2011 – 2023 which includes the 10 original SHINE challenge events
- **30 non-event** periods between 2012 – 2023
  - Eruption followed by no enhancement in proton flux
  - Eruption followed by a small or below threshold enhancement of proton flux
  - Both SOHO and GOES data sets were checked for proton enhancements since GOES has a high instrumental background
- The event and non-event lists were selected to have similar distributions of flare and CME parameters, although this was not fully achieved
- Events in both sets occurred throughout Solar Cycles 24 & 25

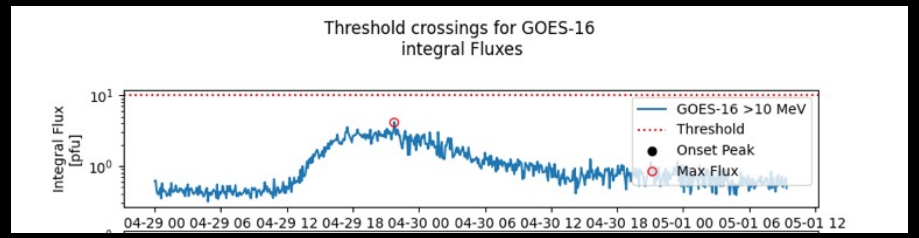
SEP Event 2011-03-08



Non-Event 2013-05-27



Non-Event 2022-04-29



SEPVAL Challenge Event Periods

SEP Events

	Original SHINE Challenge Events
SHINE	2012-03-07
SHINE	2012-03-07
SHINE	2012-05-17
SHINE	2012-07-12
SHINE	2013-04-11
SHINE	2014-01-06
SHINE	2014-01-07
SHINE	2017-07-14
SHINE	2017-09-04
SHINE	2017-09-06
SHINE	2017-09-10

	New events in Solar Cycle 25
SC25	2021-05-29
SC25	2021-10-28
SC25	2022-01-20
SC25	2022-03-28
SC25	2022-04-02
SC25	2022-08-27
SC25	2023-02-25

	Added Events in Solar Cycle 24
SC24	2011-03-08
SC24	2011-06-07
SC24	2011-08-04
SC24	2011-08-09
SC24	2012-01-23
SC24	2012-01-27
SC24	2012-03-13
SC24	2012-07-07
SC24	2012-07-23
SC24	2012-09-28
SC24	2013-05-22
SC24	2013-09-30
SC24	2014-02-25
SC24	2014-04-18
SC24	2014-09-11
SC24	2015-10-29

Event list: https://docs.google.com/spreadsheets/d/11BiYbBALP-x0n4qxURo_78rCe0bDC6kKxuEmDO2noUk/edit?usp=sharing

Non-Events

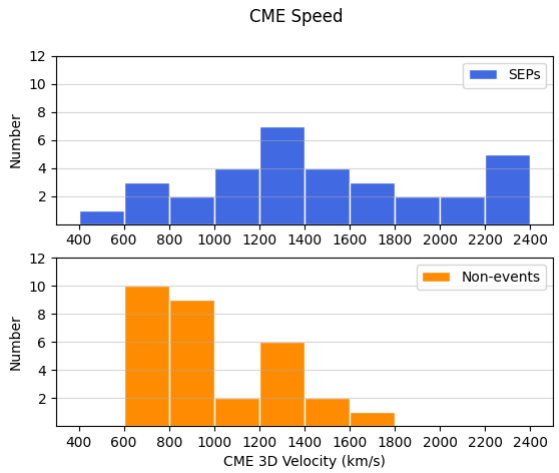
	Original SHINE Non-events
SHINE	2012-06-13
SHINE	2013-06-07
SHINE	2014-08-01
SHINE	2014-10-24
SHINE	2014-11-06
SHINE	2014-11-07
SHINE	2014-12-17
SHINE	2014-12-18
SHINE	2015-03-09
SHINE	2016-07-23
SHINE	2021-11-01
SHINE	2021-11-02
SHINE	2022-01-18
SHINE	2022-04-17

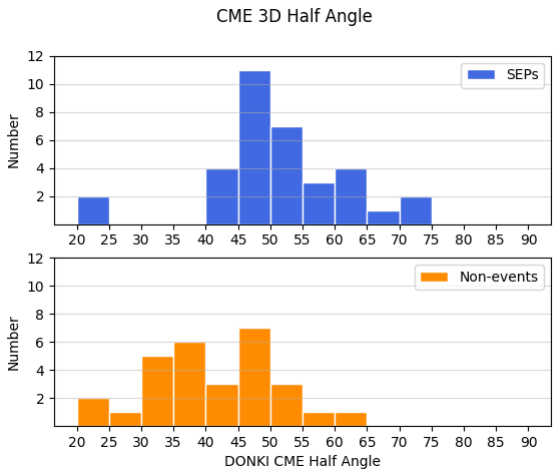
Non-Event list: https://docs.google.com/spreadsheets/d/1SPyMLBuopTp5IkMRjEGAMXVzGwcjO8sulpw7JHR78q0/edit?usp=sharing

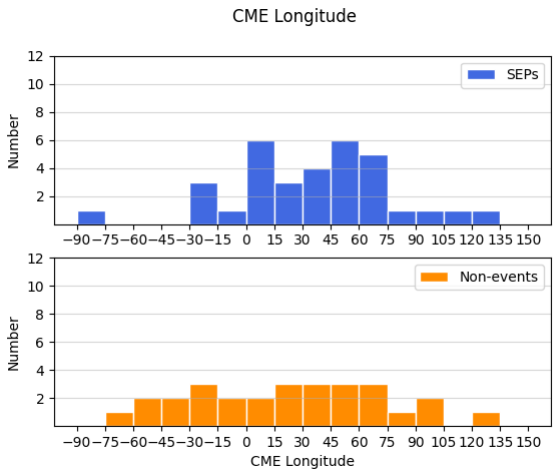
	Solar Cycle 25 Non-events
SC25	2022-04-20
SC25	2022-04-29
SC25	2022-05-25
SC25	2022-08-17
SC25	2022-08-18
SC25	2022-08-19
SC25	2022-08-29
SC25	2022-08-30
SC25	2022-12-01
SC25	2023-03-04
SC25	2023-03-06

	Added Non-events in Solar Cycle 24
SC24	2011-05-09
SC24	2012-03-04
SC24	2012-03-05
SC24	2012-06-29
SC24	2013-06-28

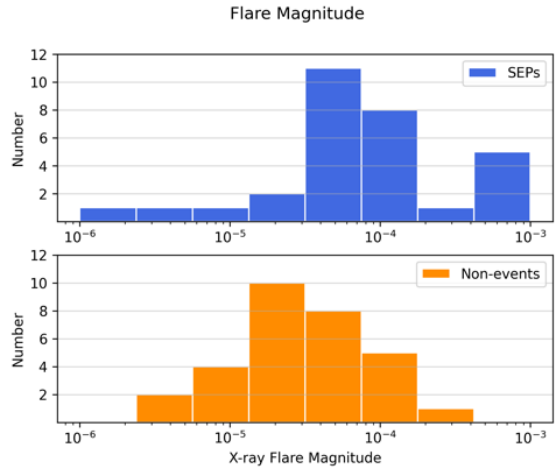
Flare and CME Distributions for the Challenge SEP Events and Non-Events

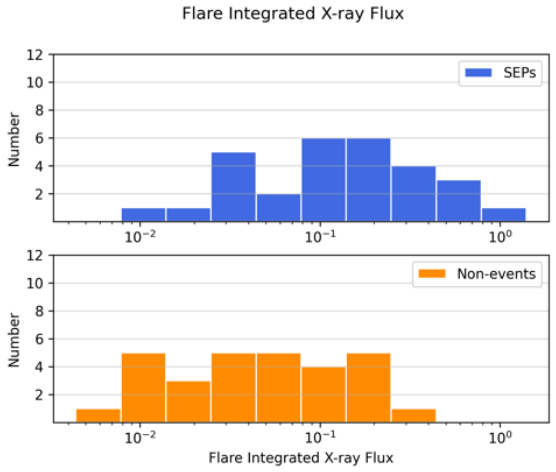


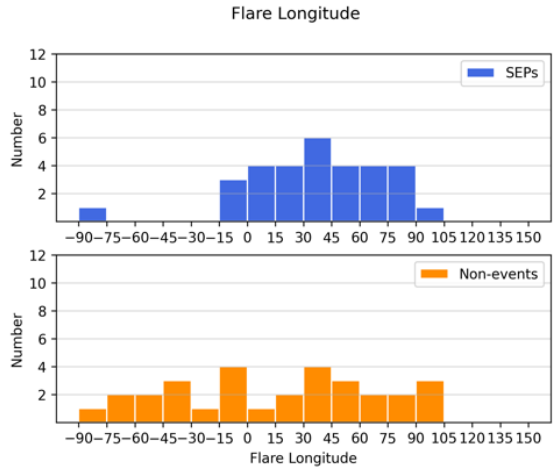


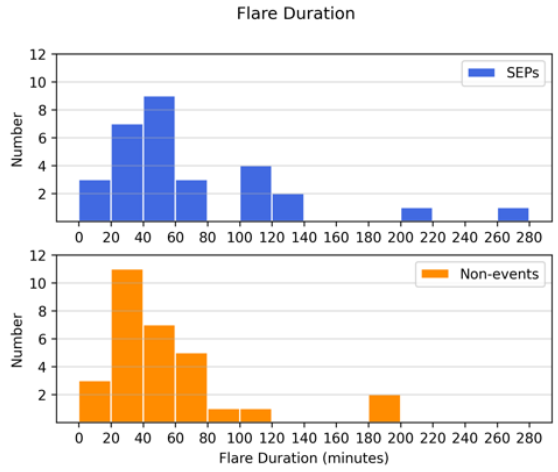


CME Parameters (left)
Flare Parameters (below)



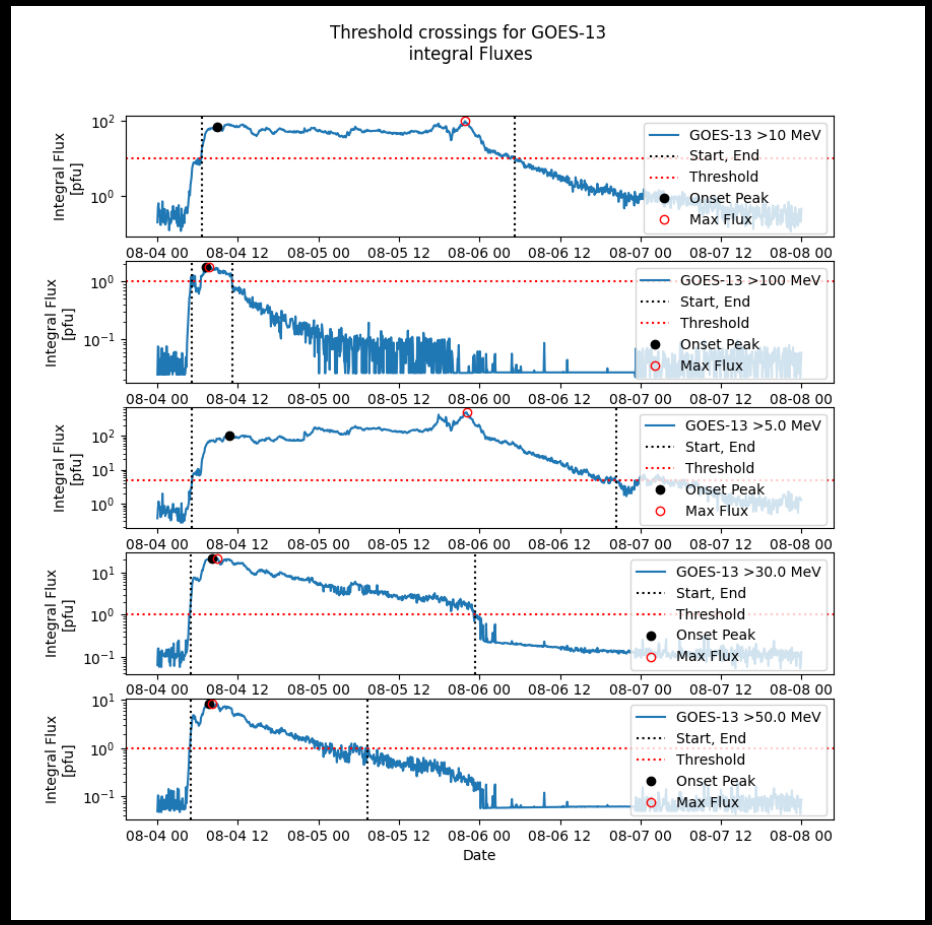






Preparation of Observations

- Used the fetchsep package developed by Katie Whitman to prepare the observational SEP event properties
  - https://github.com/ktindiana/fetchsep
- Energy channel and threshold combinations applied:
  - >5 MeV, 5 pfu
  - **>10 MeV, 10 pfu (SWPC and SRAG operational threshold)**
  - >30 MeV, 1 pfu
  - >50 MeV, 1 pfu
  - **>100 MeV, 1 pfu (SRAG operational threshold)**

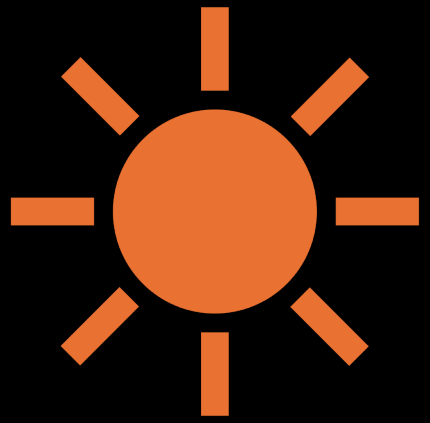


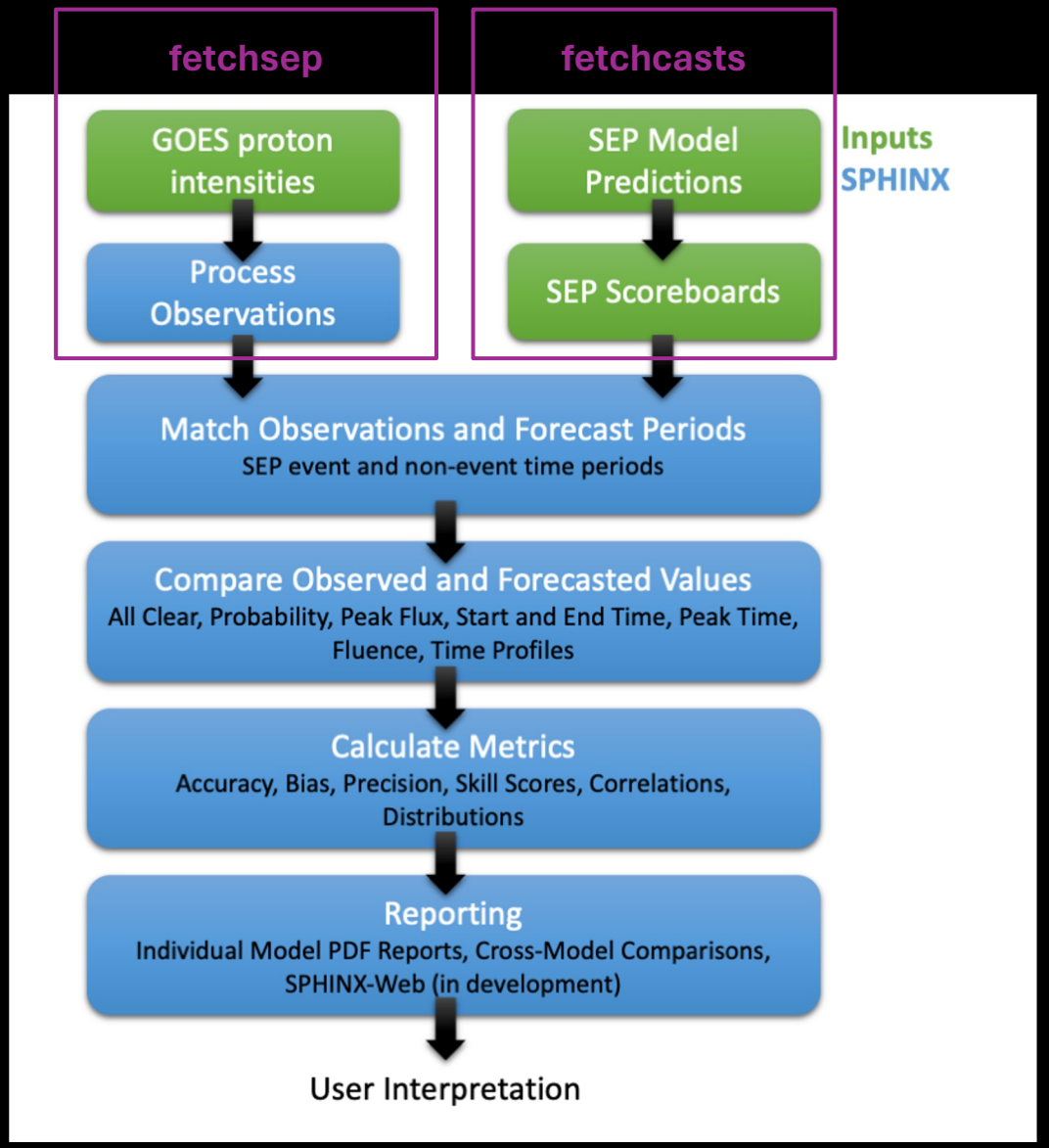
SPHINX Validation Code

Solar Particles in the Heliosphere validation INfrastructure for SpWx (SPHINX)

Goal: A generalized, automated tool that can validate any kind of forecasted quantity from any type of solar energetic particle (SEP) prediction model.

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





Validation In Visually Interactive Displays (VIVID)

VIVID

Web application for displaying the validation results of SPHINX in a dashboard of interactive plots and tables

Filter

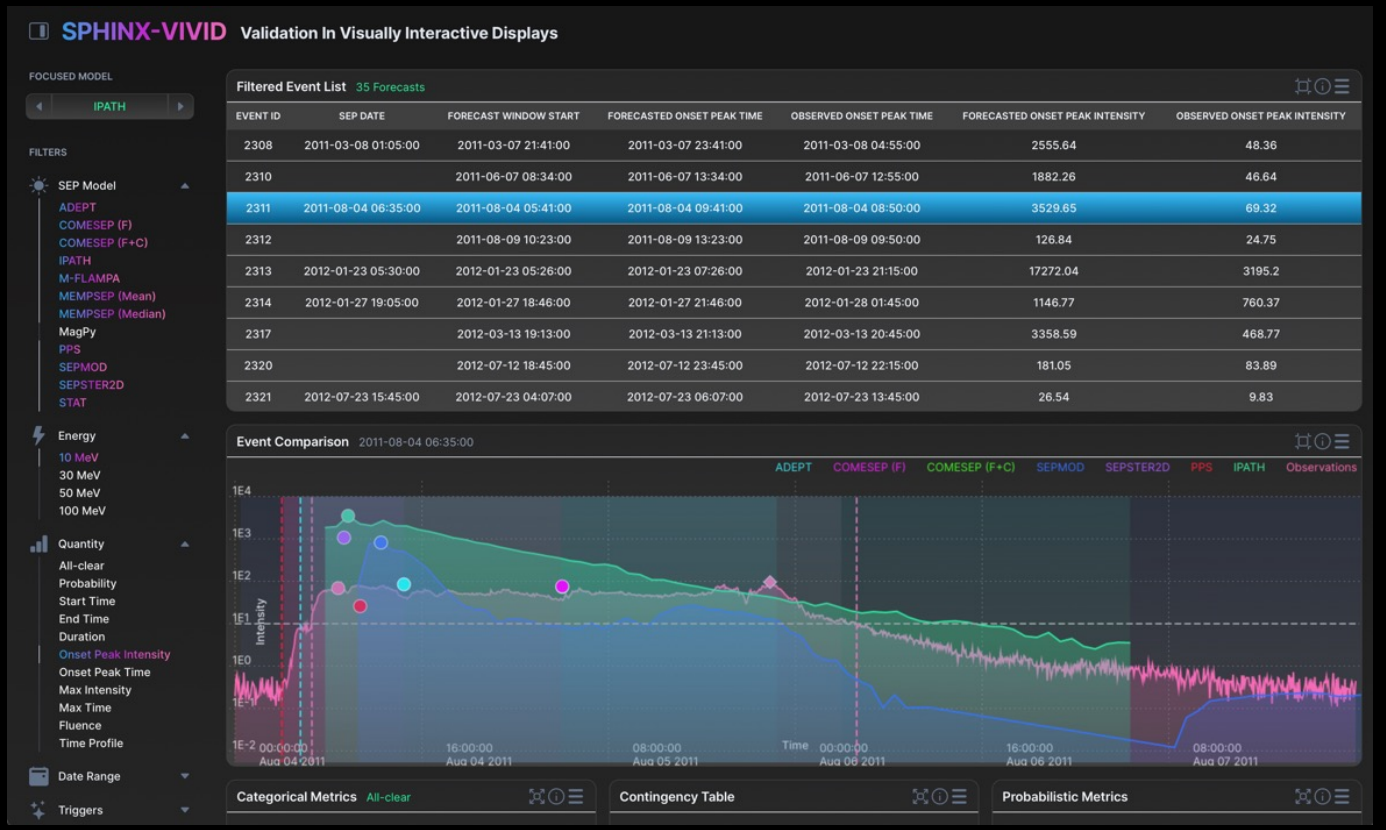
Filter results by SEP model, energy, quantity, date range, or model input – all metrics are recalculated for filtered results

Compare

Compare models side-by-side to find the state-of-the-art overall or given specific model input

Download

Download data and images for use in publications



Find poster during Thu/Fri poster session for more info

SEPVAL 2023 Model Participation

1. How well do models perform following their default workflows? Performance in a simulated real time environment.
2. What is a model's best performance? Produce predictions with no restrictions. Turning all the knobs, is it possible to get it right?
3. How do models perform with respect to physical parameters, such as eruption location, CME speed, etc? Filters applied to determine the strengths and weaknesses of our current predictive capabilities.

Question 1: How well can a model do with its default workflows? (Simulated Real Time Forecasts)	Question 2: What is the best a model can do? (Science)
ADEPT AFRL PPS ASPECS COMESEP (flare, flare + CME) iPATH Lavasa Model MEMPSEP (Mean, Median, 10 submodels) MagPy REleASE SEPMOD SEPSTER SEPSTER2D SPREAdFAST UMASEP Zhang et al. (cRT, rRT, SEPSAT)	M-FLAMPA STAT

Forecasts submitted to SEPVAL 2023

Model	Developer Point of Contact	Affiliation	Energy Channels	Forecasted Quantities	# Forecasts Submitted	# Forecasts Processed
ADEPT 1hr, 6hr	Stephen White	US Air Force	>10 MeV	Time Profile	25	25
COMESEP flare, flare+CME	Mark Dierckxsens	BIRA	>10	Probability, Peak	63, 63	60, 63
cRT+AE10	Ming Zhang	Florida Institute of Technology	>10	Probability	63	63
ENLIL+SEPMOD	Janet Luhmann	UC Berkeley	>10, >30, >50, >100	Time Profile	61	61
Lavasa	Eleni Lavasa	National Observatory of Athens	>10	All Clear	58	58
MagPy	David Falconer, Tilaye Tedesse	UA Huntsville, NASA JSC SRAG	>10	Probability	2182	2182
MEMPSEP Mean, Median	Subhamoy Chatterjee	Southwest Research Institute	>10	Probability	60, 60	60, 60
MFLAMPA	Igor Sokolov	University of Michigan	>10, >30, >50, >100	Time Profile	9	8
PPS (SFS Update)	Stephen White	US Air Force	>10, >100	Peak Flux	61	61
SEPSAT	Ming Zhang	Florida Institute of Technology	>10, >100	Time Profile	64	64
SEPSTER	Ian Richardson	University of Maryland	>10, >30, >50, >100	Peak Flux	64	64
SEPSTER2D	Alessandro Bruno	NASA GSFC	>10, >30, >50, >100	Peak, Fluence	60	60
SPREAdFAST	Kamen Kozarev	Bulgarian Academy of Sciences	>10, >30, >50, >100	Time Profile	8	8
SPRINTS 0-24, 24-48, 48-72, 72-96	Alec Engell	NextGen	>10, >30, >50, >100	Probability, Peak	13134, 13134, 13134, 13134	13134, 13134, 13134, 13134
STAT	Jon Linker	Predictive Science, LLC	>10, >30, >50, >100	Time Profile	6	6
UMASEP-10, -100	Marlon Nunez	University of Malaga	>10, >30, >50, >100	Peak, Start	27572, 32240	27161, 32106
UNSPELL	Sigiava Alminalragia-Giamini	SPARC	>10	Probability	61	61
ZEUS+iPATH	Gang Li, Junxiang Hu	UA Hunstville, NASA GSFC	>10, >30, >50, >100	Time Profile	57	56

SAWS-ASPECS

Developer: Athanasios Papaioannou+, Affiliation: National Observatory of Athens, Energy Channels: >10, >100 MeV

Model SAWS-ASPECS	Forecasted Quantities	# Forecasts Submitted	# Forecasts Processed
CME (CACTus) 50%	Time Profile	60	60
CME (CACTus) 90%	Time Profile	60	60
CME (CACTus)	Probability/All Clear	60	60
CME (CDAW) 50%	Time Profile	63	63
CME (CDAW) 90%	Time Profile	63	63
CME (CDAW)	Probability/All Clear	63	63
Flare+CME (CACTus) 50%	Time Profile	57	57
Flare+CME (CACTus) 90%	Time Profile	57	57
Flare+CME (CACTus)	Probability/All Clear	57	57
Flare+CME (CDAW) 50%	Time Profile	60	60
Flare+CME (CDAW) 90%	Time Profile	60	60
Flare+CME (CDAW)	Probability/All Clear	60	60
Flare 50%	Time Profile	60	60
Flare 90%	Time Profile	60	60
Flare	Probability/All Clear	60	60

Model SAWS-ASPECS	Forecasted Quantities	# Forecasts Submitted	# Forecasts Processed
CME (CACTus) electrons 50%	Time Profile	60	60
CME (CACTus) electrons 90%	Time Profile	60	60
CME (CACTus) electrons	Probability/All Clear	60	60
CME (CDAW) electrons 50%	Time Profile	63	63
CME (CDAW) electrons 90%	Time Profile	63	63
CME (CDAW) electrons	Probability/All Clear	63	63
Flare+CME (CACTus) electrons 50%	Time Profile	57	57
Flare+CME (CACTus) electrons 90%	Time Profile	57	57
Flare+CME (CACTus) electrons	Probability/All Clear	57	57
Flare+CME (CDAW) electrons 50%	Time Profile	60	60
Flare+CME (CDAW) electrons 90%	Time Profile	60	60
Flare+CME (CDAW) electrons	Probability/All Clear	60	60
Flare electrons 50%	Time Profile	60	60
Flare electrons 90%	Time Profile	60	60
Flare electrons	Probability/All Clear	60	60

Quantities Validated by SPHINX

All forecasting fields that can be submitted to the SEP Scoreboards are validated by SPHINX.

Will an SPE occur?

- All Clear (threshold crossed/not crossed)
- Probability of Occurrence

When will it happen? How long will it last?

- Start Time
- End Time
- Duration
- Peak Time

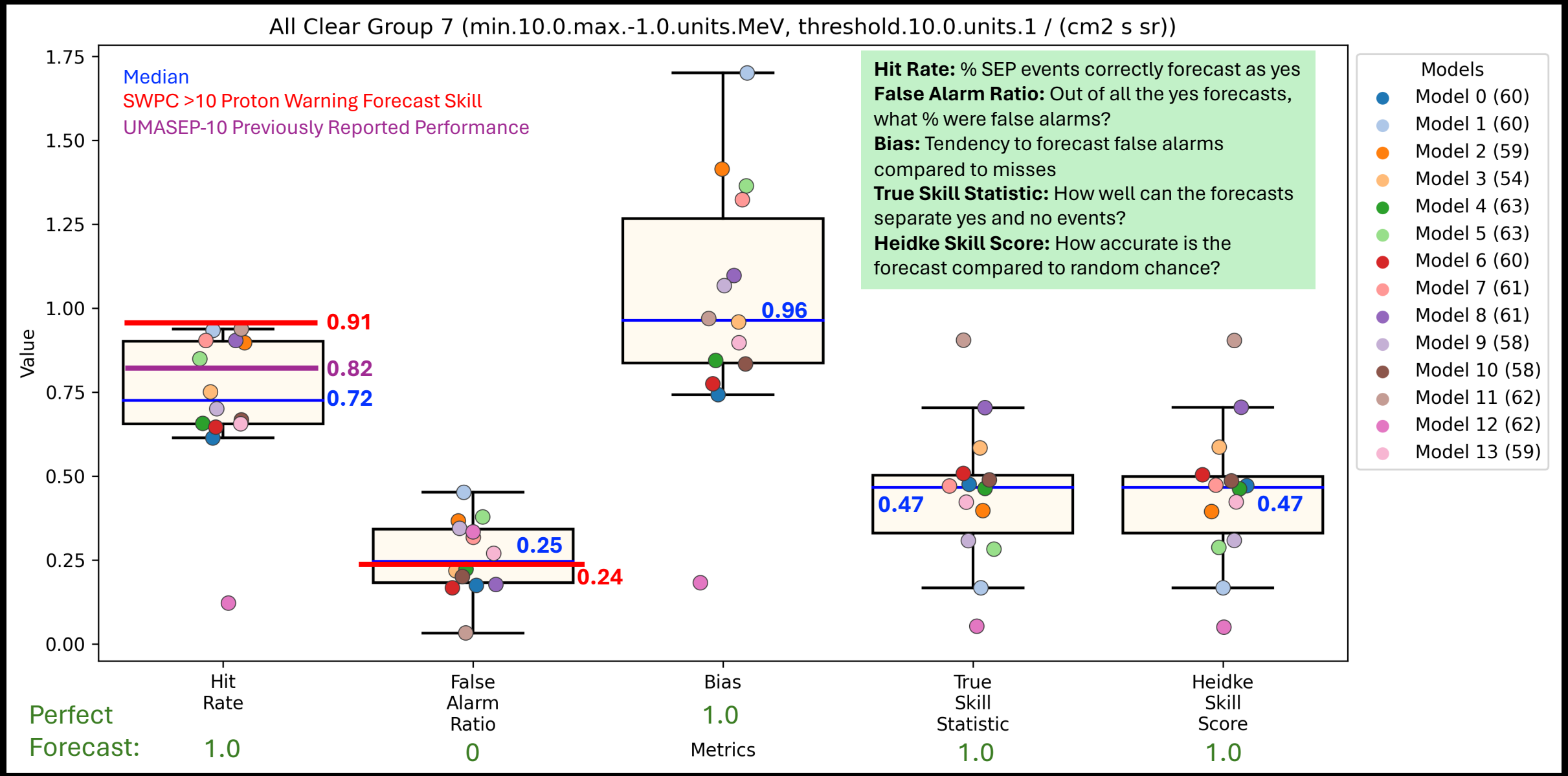
How big will it be?

- Peak Flux
- Fluence
- SEP Intensity Time Profile

How much warning did we get?

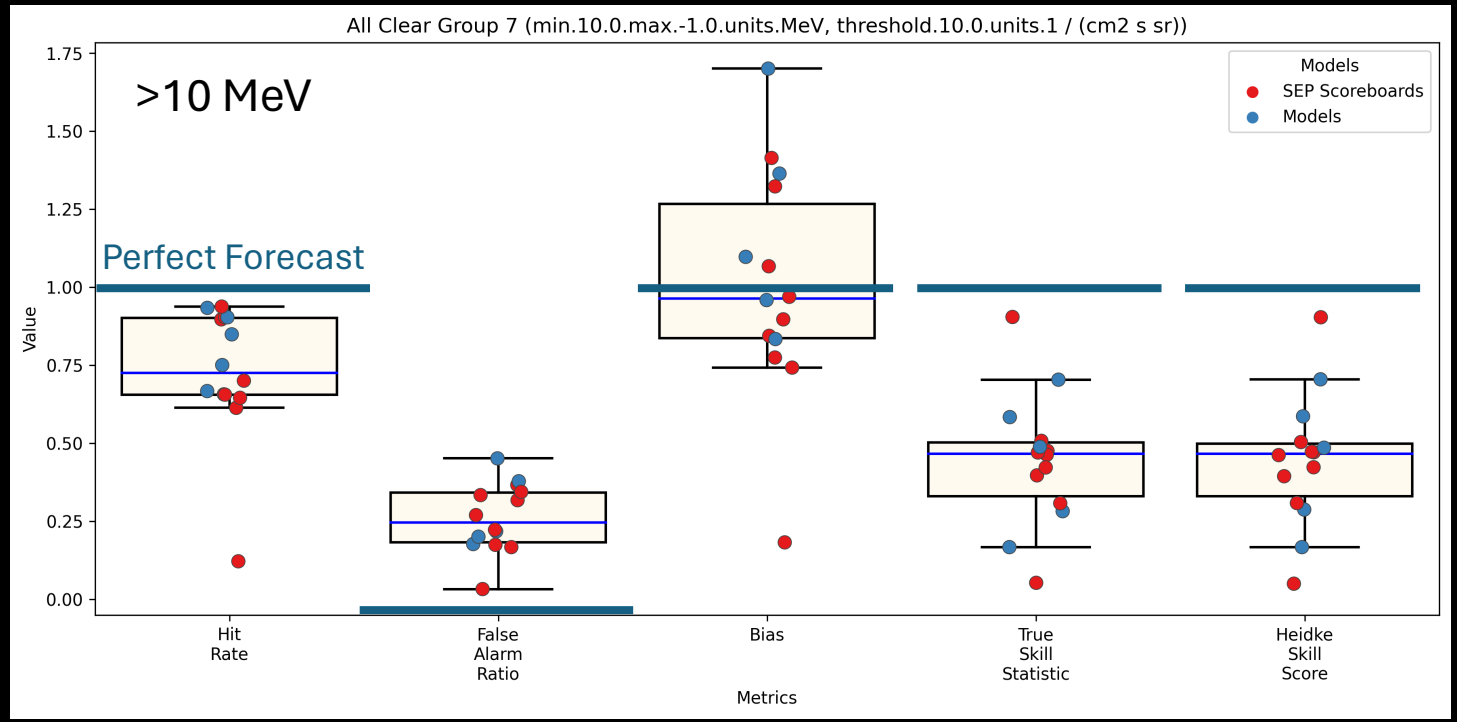
- Advanced Warning Time

Example All Clear Performance (>10 MeV)



Performance of SEP Scoreboard Models in the SEPVAL Challenge

- The SEP Scoreboards contain models across all levels of performance, including top performers
- It is a resource where we are trying out different forecasting techniques
- Validation is starting to show which approaches are more and less effective
- Identify models that have the potential to transition to higher Readiness Levels (e.g. for selection by SWPC)
- Identify models that may need additional work to improve performance
- **All interested developers are encouraged to participate in the SEP Scoreboards**



Activities Enabled by the SEPVAL Community Challenges

- Designed, thorough, structured challenge gave an independent characterization of model performance
- Provided an understanding of the state-of-the-art (SOA) performance of SEP models across the research community (never been done before)
- Receiving output from so many types of models enabled the development of SPHINX as a generalized tool – exercised failure space, tested logic
- SEPVAL Contributors are now better-prepared to participate in the SEP Scoreboards – familiar with JSON format
- We can establish target benchmarks for model performance now that we have a clearer view of SOA
- We can identify which metrics are most meaningful for SRAG (and others)

Activities Enabled by the SPHINX Framework

- Perform an independent verification of self-reported metrics
- Establish new benchmarks
- Provide feedback to SEP model developers to inspire improvements in an R2O2R cycle
- Identify which aspects of performance a model needs to improve to reach benchmark goals
- Readily reassess performance after model updates
- Provide a tool to evaluate gap closure
- EOY: Continuous real time assessment of SEP Scoreboards performance with SPHINX (to become public via CCMC)

Data Sampling affects Validation Outcomes:
Community Challenges and SEP Scoreboards

- Validation through the community challenges and for forecasts in the SEP Scoreboards tells different but complimentary stories
- We are working to understand how to best interpret each type of validation
- The SEPVAL community challenge provides a broad view of SEP forecasting skill across the research field and the varying approaches that are being developed and tested
- The SEP Scoreboards allow us to directly evaluate real time performance

- **The community challenge** using a benchmark event list provides information about (but not limited to):
  - Model performance relative to other participating models
  - Which model approaches are doing comparatively well
  - Which model approaches may need improvement

- **SEP Scoreboard real time validation** gives information about (but not limited to) how models perform in a real time setting:
  - All false alarms and misses
  - Impact of input data availability
  - Advanced warning time
  - Timing of overall operational forecasting workflows

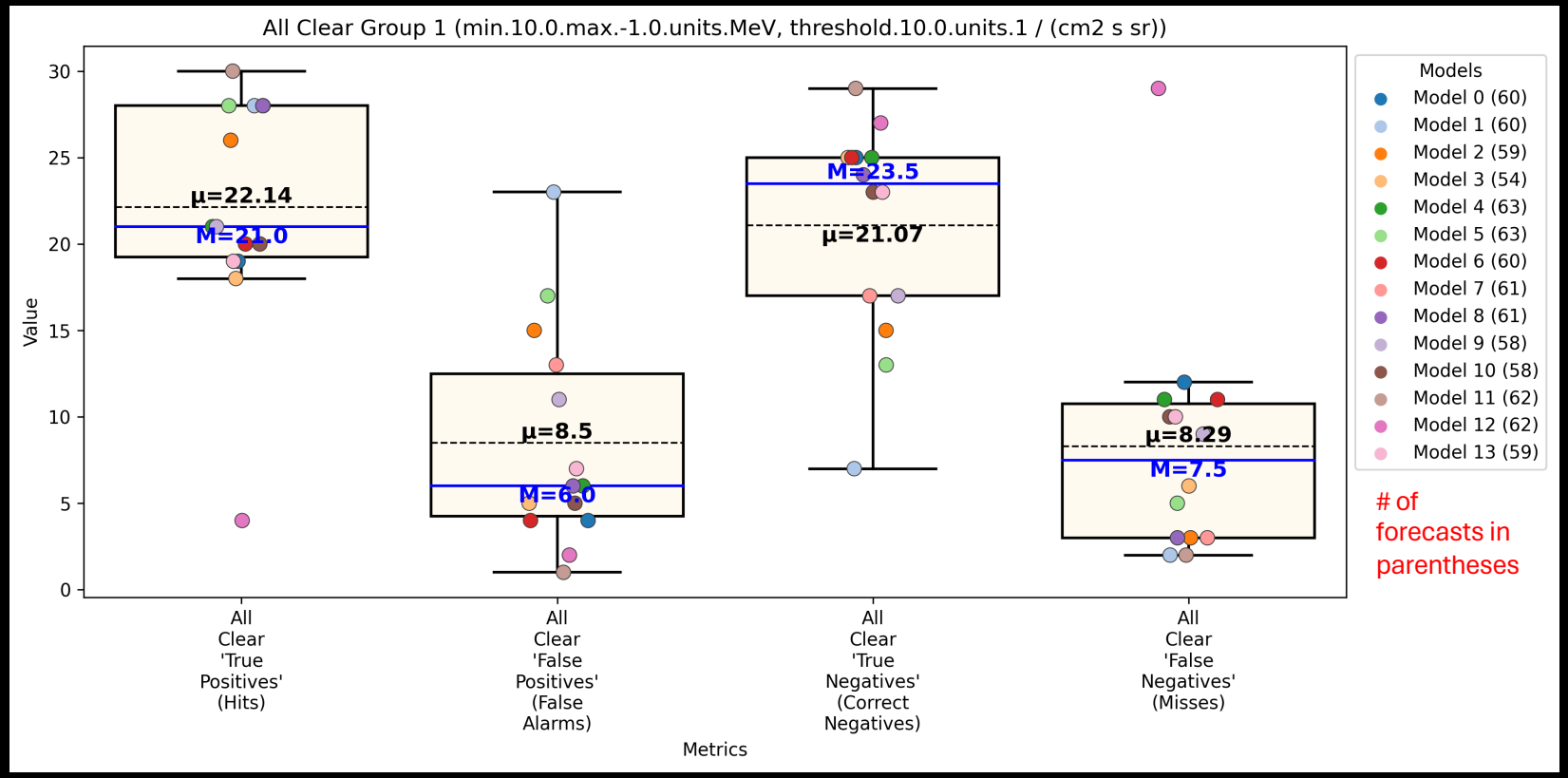
Your Feedback is Welcome

- The results presented here are a subset of the information generated by SPHINX.
- Each model developer that contributed to the SEPVAL 2023 challenge received an individual validation report and supporting summary plots with a full set of metrics.
- The SEPVAL 2023 results will be reported in 2024 at TESS, EGU, COSPAR and ESWW using anonymized plots.
- Model developers are encouraged to contact Katie Whitman and the SPHINX development team to discuss their personalized results, ask any questions, for clarifications, or to motivate revisions in the SPHINX logical workflow.
- This is a collaborative R2O2R effort and the ISEP project and the development of SPHINX has benefitted immensely from the participation of the community.
- A publication will be prepared. We will be contacting model developers to participate in writing the paper and agreeing upon the results.

Further Examples of Metrics

The models have been anonymized until the metrics have been discussed with the model developers and agreed upon for public release

SEPVAL 2023: All Clear Summary



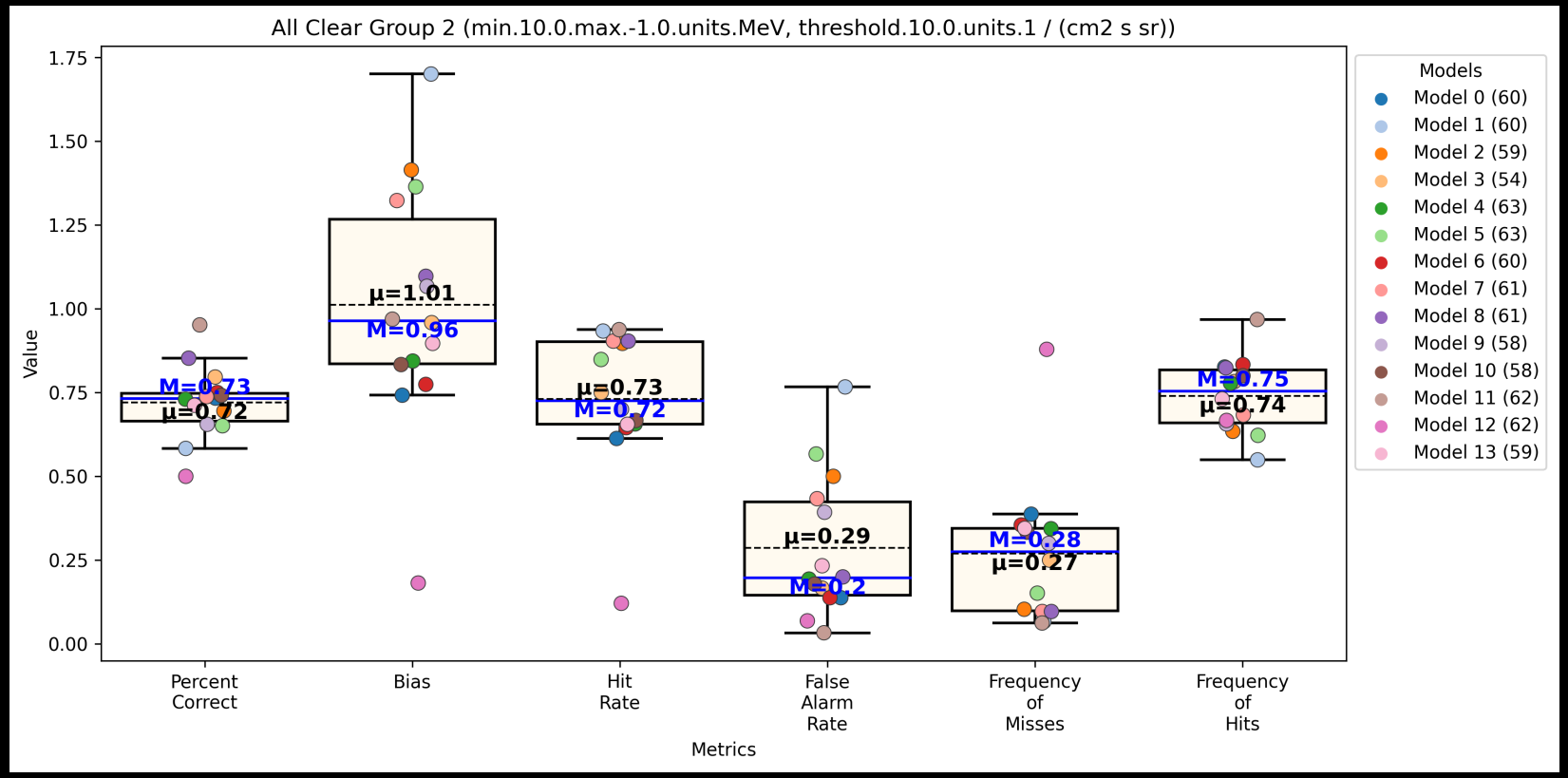
Perfect Forecast:	33	0	30	0

Models with multiple forecasts per time period: MagPy, SPRINTS, and UMASEP produce new forecasts each hour, 3 minutes, or 1 minute (respectively).

The metrics presented here have been reduced to an overall forecast result per challenge event time period.

Hit: At least one forecast indicating an SEP event was issued for an observed SEP
Miss: All forecasts were All Clear but an SEP was observed
False Alarm: At least one Not Clear forecast was issued during a non-event period
Correct Negative: All forecasts were All Clear for a non-event period

SEPVAL 2023: All Clear Summary >10 MeV, 10 pfu



	Percent Correct	Bias	Hit Rate	False Alarm Rate	Frequency of Misses	Frequency of Hits
Perfect Forecast:	1.0	1.0	1.0	0	0	1.0

H – Hits
M – Misses
FA – False Alarms
CN – Correct Negatives
N – Total forecasts

Percent Correct:

$$\frac{H + CN}{N}$$

Bias:

$$\frac{H + FA}{H + M}$$

Hit Rate:

$$\frac{H}{H + M}$$

False Alarm Rate:

$$\frac{FA}{FA + CN}$$

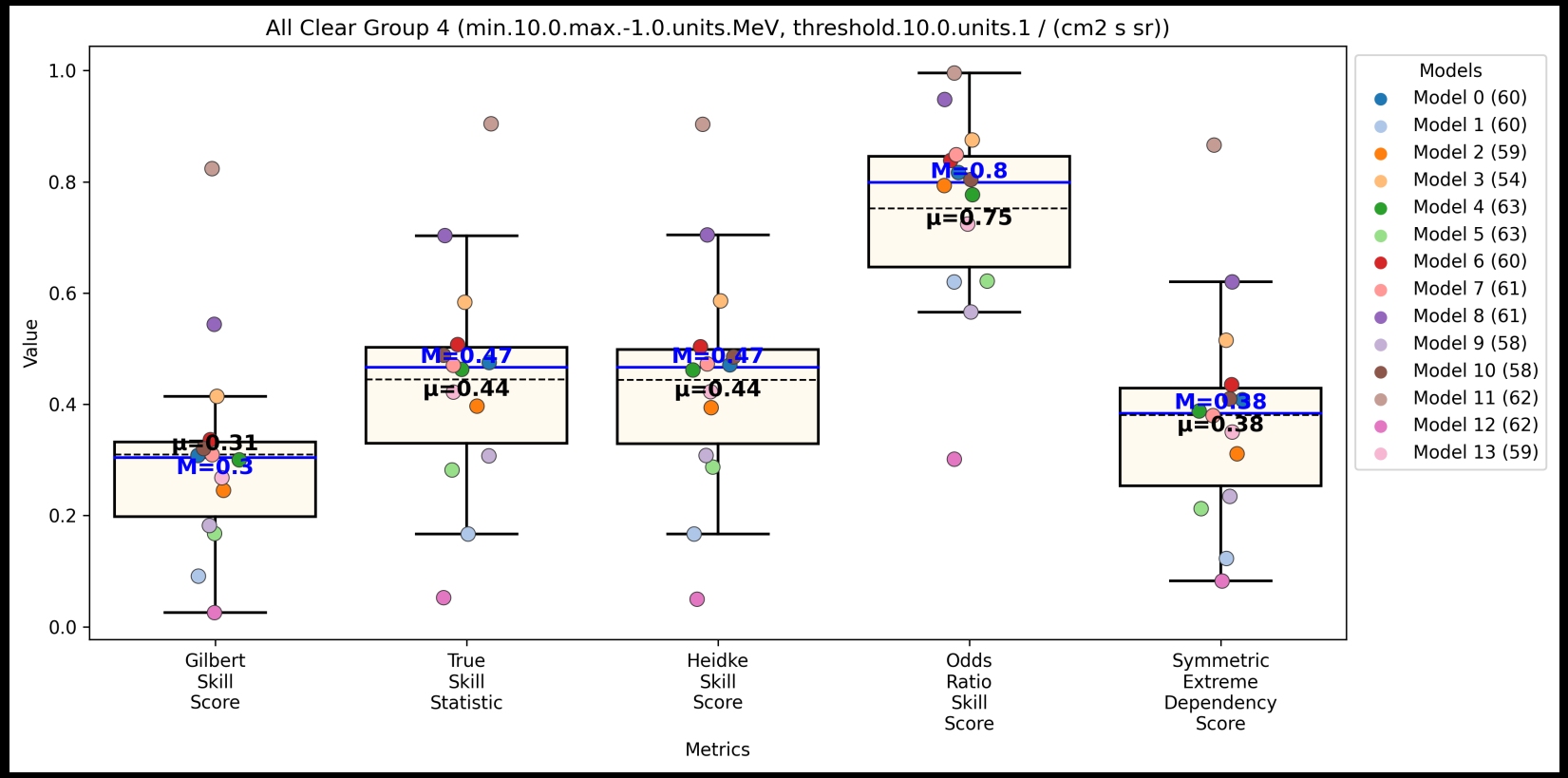
Frequency of Misses:

$$\frac{M}{H + M}$$

Frequency of Hits:

$$\frac{H}{H + FA}$$

SEPVAL 2023: All Clear Summary >10 MeV, 10 pfu



Perfect Forecast:	1.0	1.0	1.0	1.0	1.0

Gilbert Skill Score:
How well did the forecast yes events correspond to the observed yes events, accounting for hits by chance?

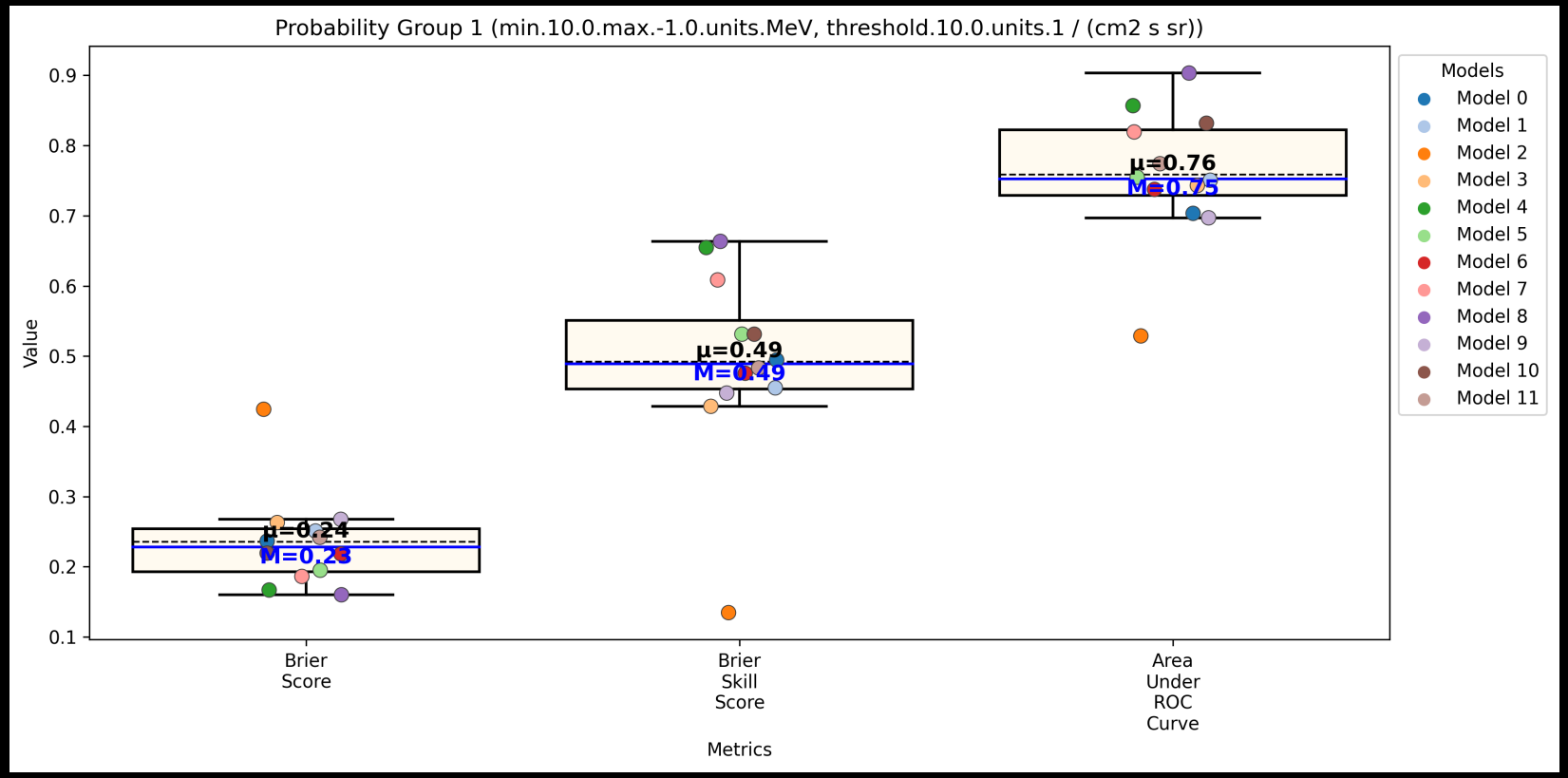
True Skill Statistic:
How well did the forecast separate the yes events from the no events?

Heidke Skill Score:
What was the accuracy of the forecast relative to random chance?

Odds Ratio Skill Score:
What was the improvement over random chance?

Symmetric Extreme Dependency Score:
Emphasizes hits and de-emphasizes correct negatives.

SEPVAL 2023: Probability Summary >10 MeV, 10 pfu



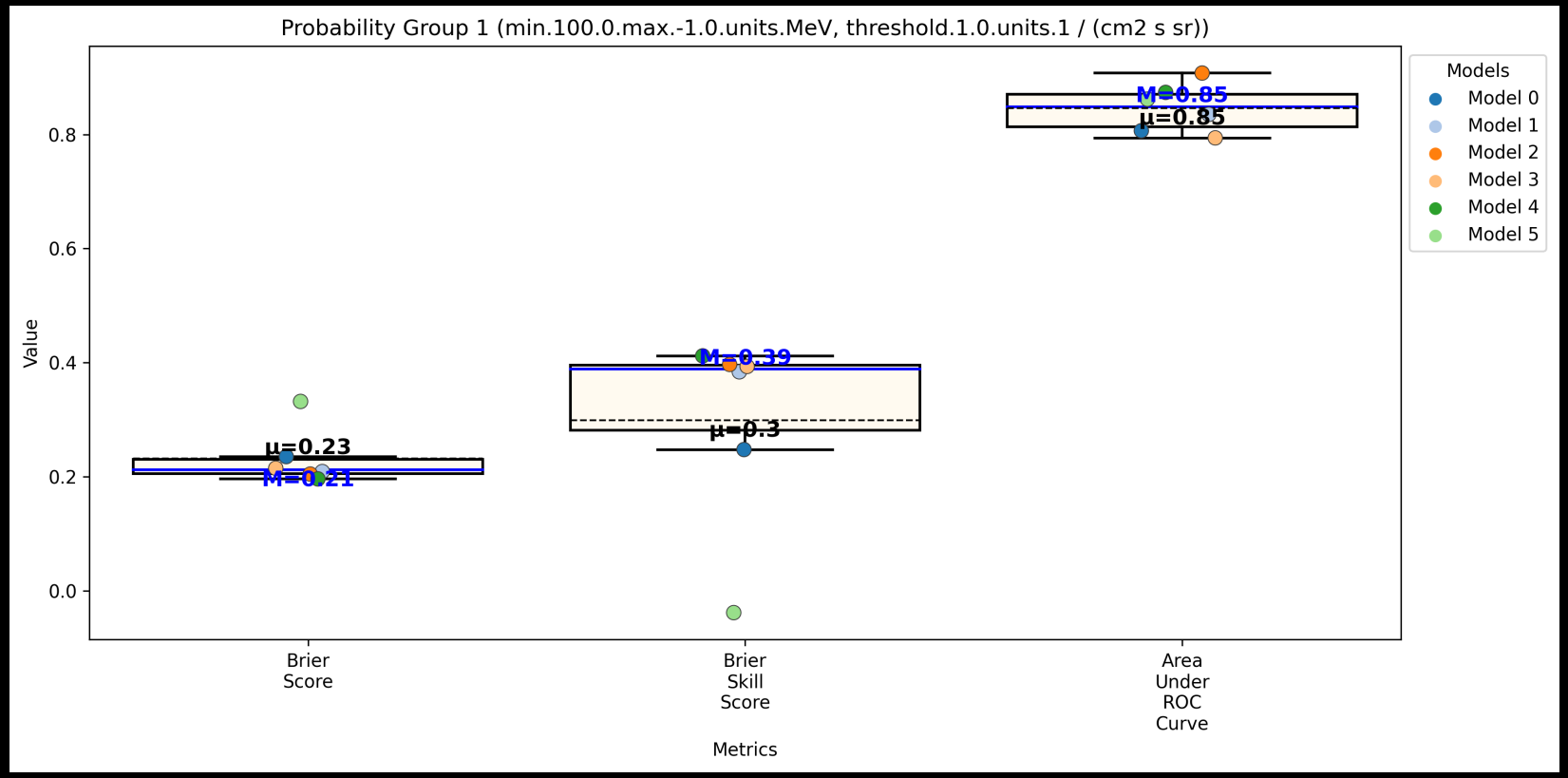
Brier Score:
$1/N \; SUM(f - o)^2$

Brier Skill Score:
Brier Score compared to climatology. Here the climatology was set to the probability value provided in Bain et al. 2021 for SWPC.
1.0 is a perfect forecast. 0.0 is the skill of the reference.

Area Under ROC Curve:
Relative value as a function of a user's cost/loss ratio. 1.0 is a perfect forecast. 0.0 indicates the value of climatology.

Perfect Forecast:	0	1.0	1.0

SEPVAL 2023: Probability Summary >100 MeV, 1 pfu



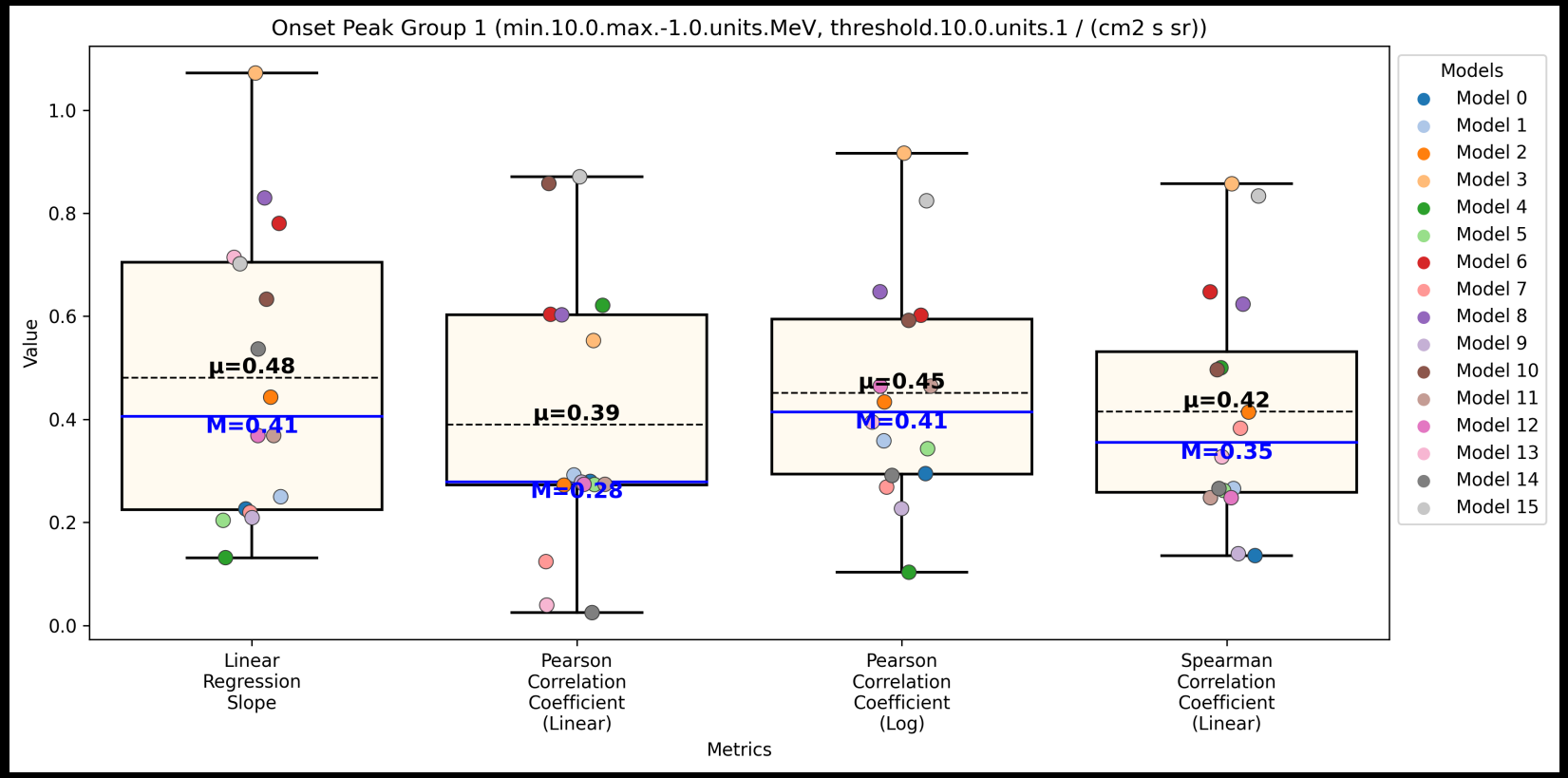
Brier Score:
$1/N \, SUM(f - o)^2$

Brier Skill Score:
Brier Score compared to climatology. Here the climatology was set to the probability value provided in Bain et al. 2021 for SWPC. 1.0 is a perfect forecast. 0.0 is the skill of the reference.

Area Under ROC Curve:
Relative value as a function of a user's cost/loss ratio. 1.0 is a perfect forecast. 0.0 indicates the value of climatology.

Perfect Forecast:	0	1.0	1.0

SEPVAL 2023: Onset Peak Summary >10 MeV, 10 pfu



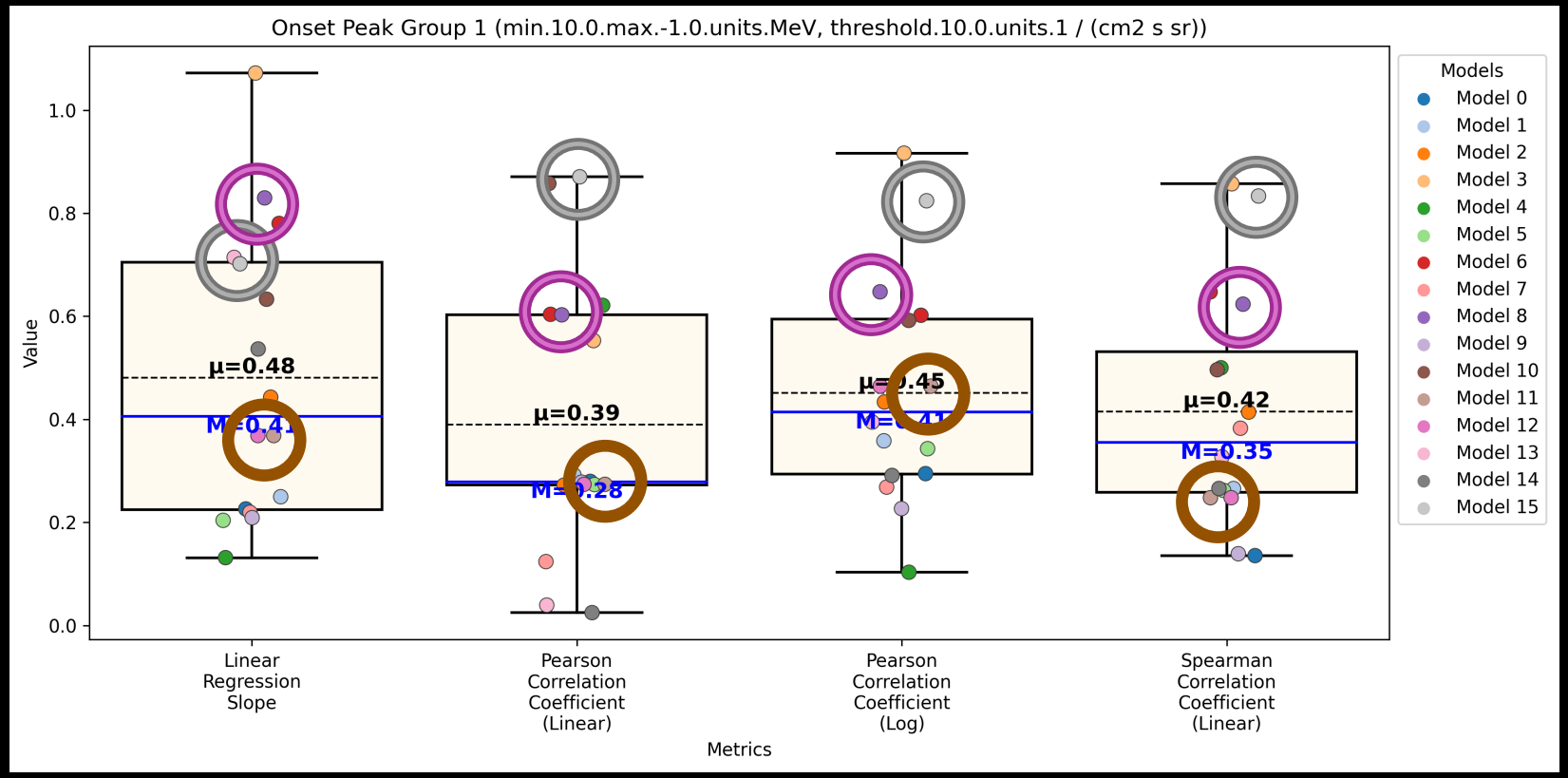
Perfect Forecast: 1.0 1.0 1.0 1.0

Linear Regression Slope:
Slope of a regression line fit to the correlated observations and forecasts.

Pearson Correlation Coefficient:
Correlation between two variables that each have normal distributions.

Spearman Correlation Coefficient:
Correlation of data by rank order rather than the actual values. Describes the monotonic relationship between observations and forecasts.

SEPVAL 2023: Onset Peak Summary >10 MeV, 10 pfu



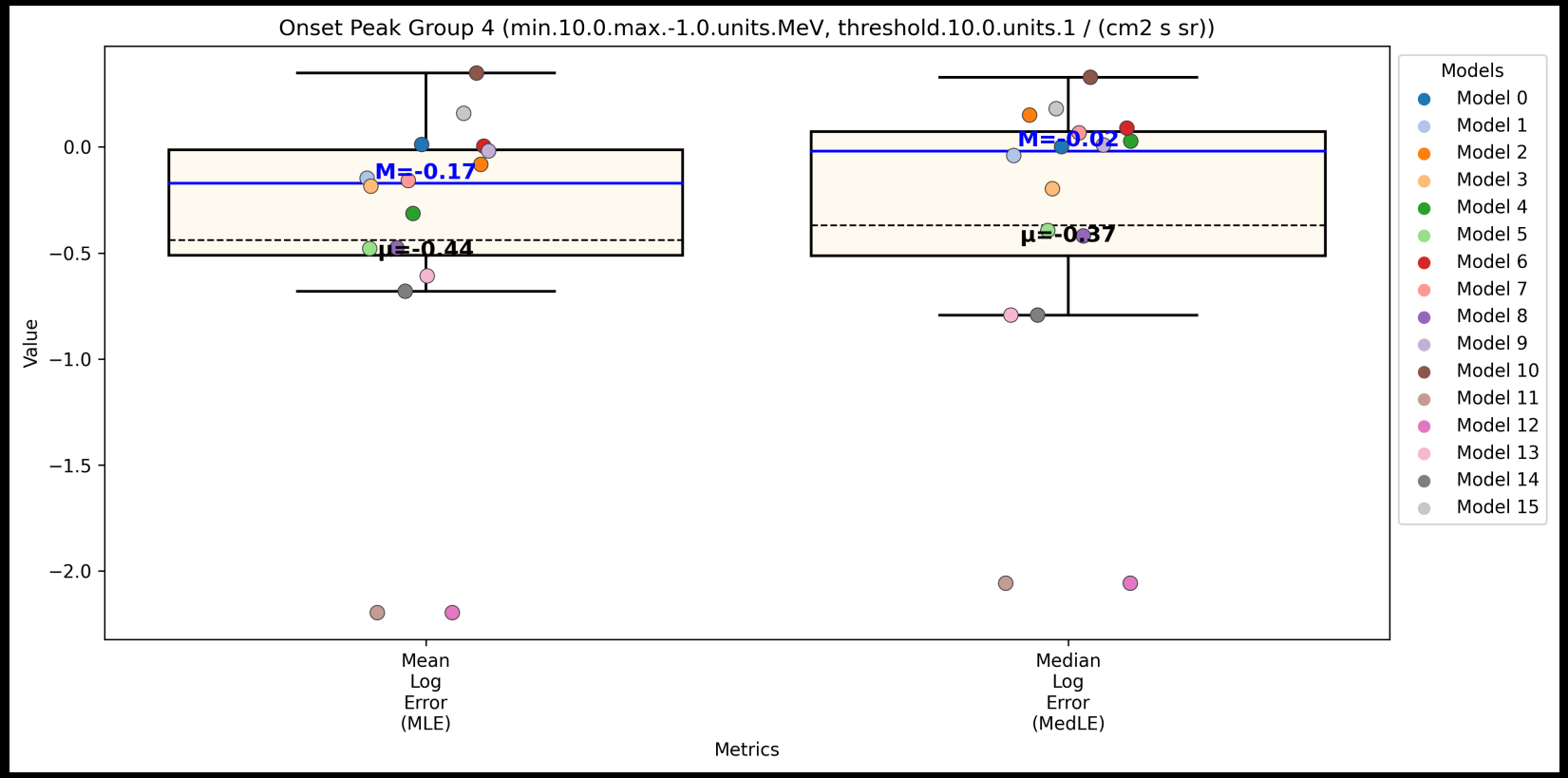
Linear Regression Slope:
Slope of a regression line fit to the correlated observations and forecasts.

Pearson Correlation Coefficient:
Correlation between two variables that each have normal distributions.

Spearman Correlation Coefficient:
Correlation of data by rank order rather than the actual values. Describes the monotonic relationship between observations and forecasts.

Perfect Forecast:	1.0	1.0	1.0	1.0

SEPVAL 2023: Onset Peak Summary >10 MeV, 10 pfu



Mean Log Error:
MEAN(log10(f) – log10(o))

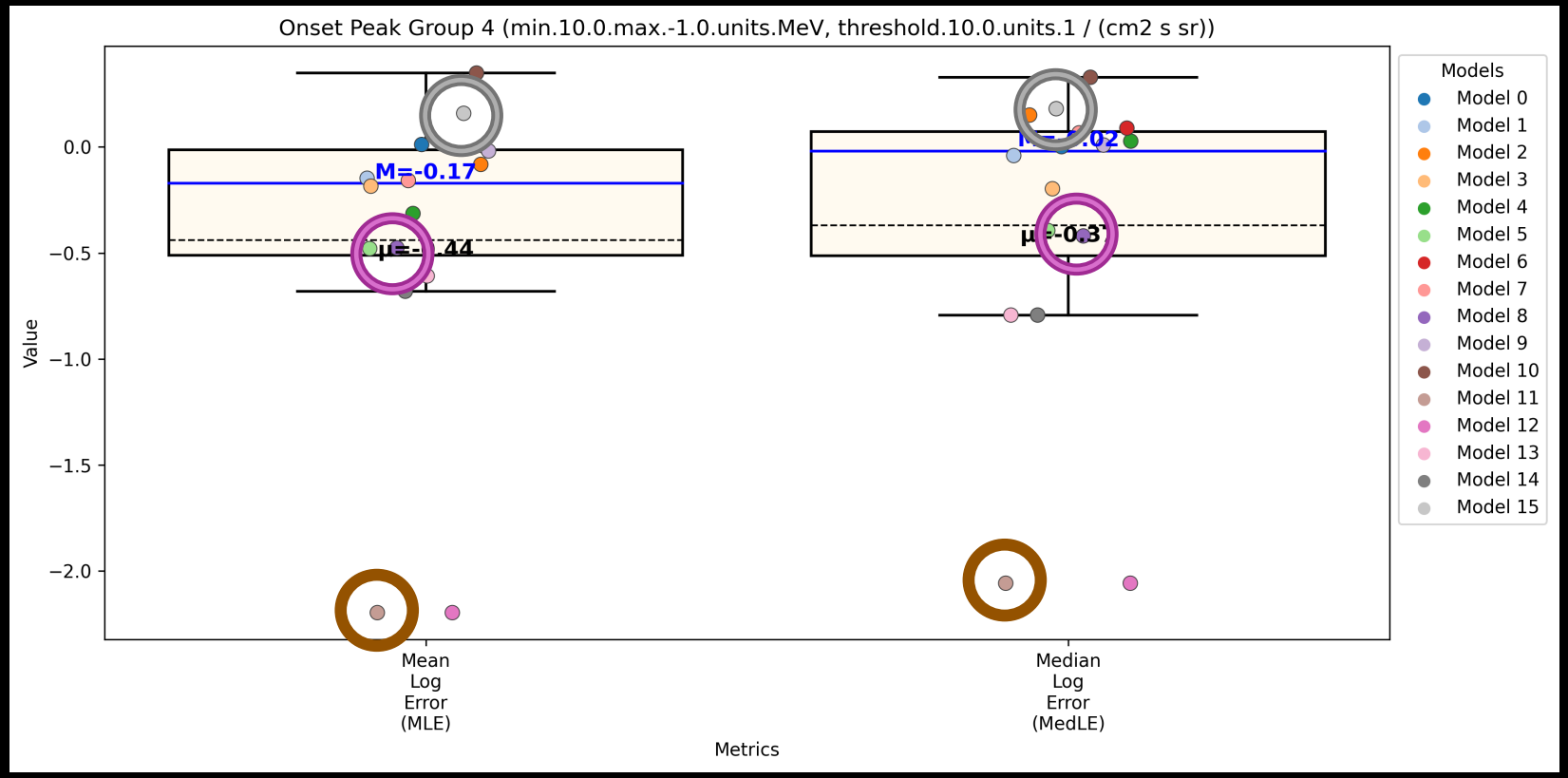
Median Log Error:
MEDIAN(log10(f) – log10(o))

MLE and MedLE are a measure of bias.

Negative = Underprediction
Positive = Overprediction

Perfect Forecast:	0	0

SEPVAL 2023: Onset Peak Summary >10 MeV, 10 pfu



Mean Log Error:
MEAN(log10(f) – log10(o))

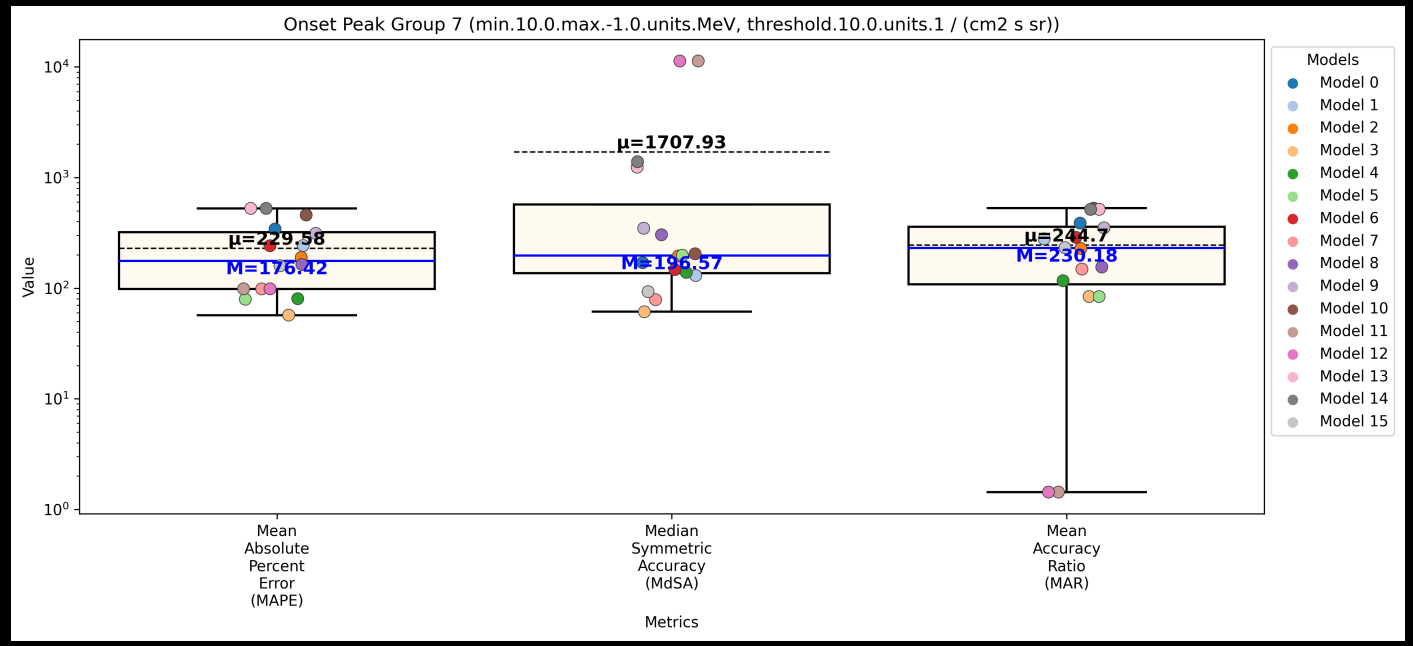
Median Log Error:
MEDIAN(log10(f) – log10(o))

MLE and MedLE are a measure of bias.

Negative = Underprediction
Positive = Overprediction

Perfect Forecast:	0	0

SEPVAL 2023: Onset Peak Summary >10 MeV, 10 pfu



Perfect Forecast: 0% 0% 100%

All metrics are in linear space.

Mean Absolute Percent Error:

$$\frac{|f - o|}{o} \times 100\%$$

Difference between the forecast and observation relative to the observed value.

Median Symmetric Accuracy:

exp(MEDIAN(|ln(o/f)|) – 1) x 100%

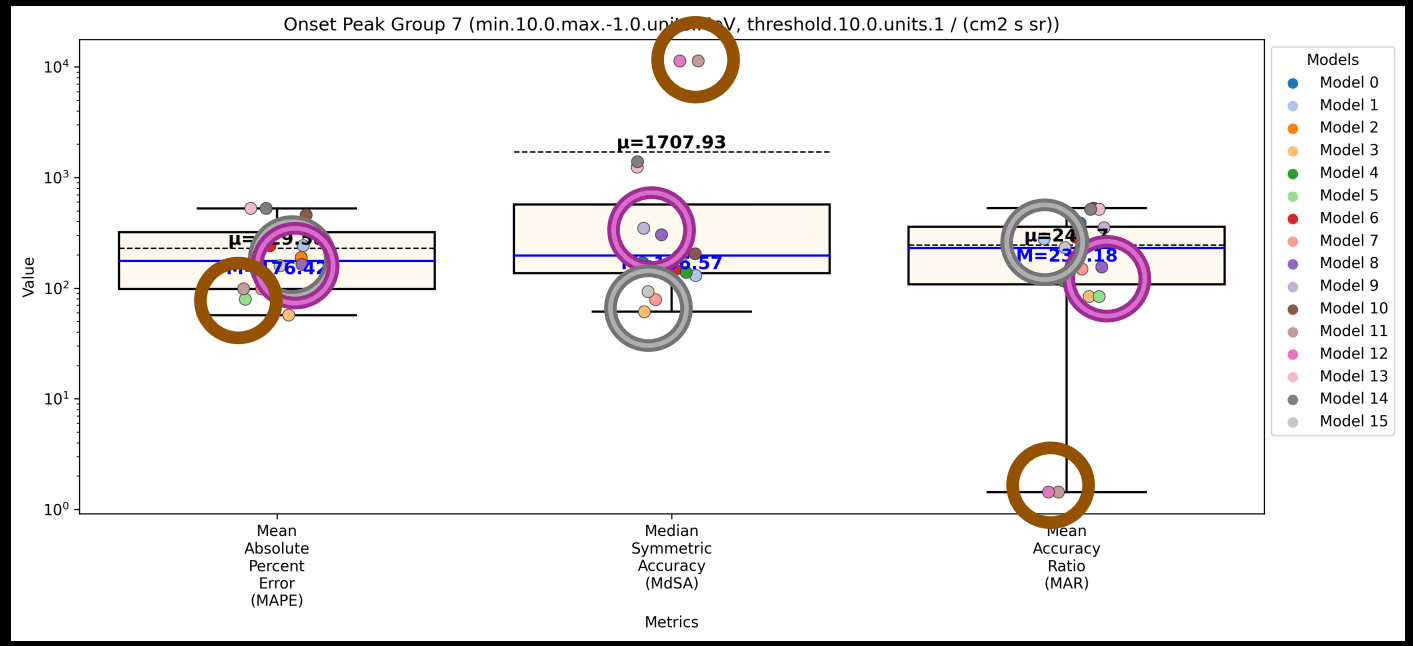
Morley et al. 2018
Penalizes over and underestimates equally.

Mean Accuracy Ratio:

$$\frac{f}{o} \times 100\%$$

< 100% = underestimate
>100% = overestimate

SEPVAL 2023: Onset Peak Summary >10 MeV, 10 pfu



All metrics are in linear space.

Mean Absolute Percent Error:

$$\frac{|f - o|}{o} \times 100\%$$

Difference between the forecast and observation relative to the observed value.

Median Symmetric Accuracy:

exp(MEDIAN(|ln(o/f)|) – 1) x 100%

Morley et al. 2018
Penalizes over and underestimates equally.

Mean Accuracy Ratio:

$$\frac{f}{o} \times 100\%$$

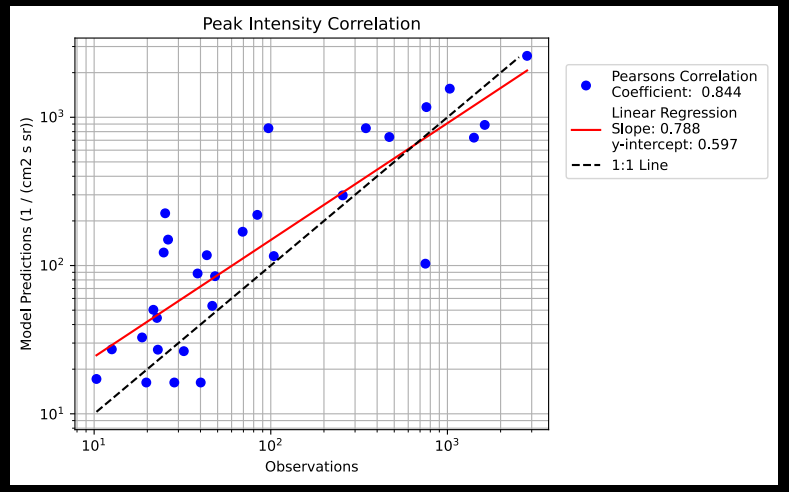
< 100% = underestimate
>100% = overestimate

Perfect Forecast:	0%	0%	100%

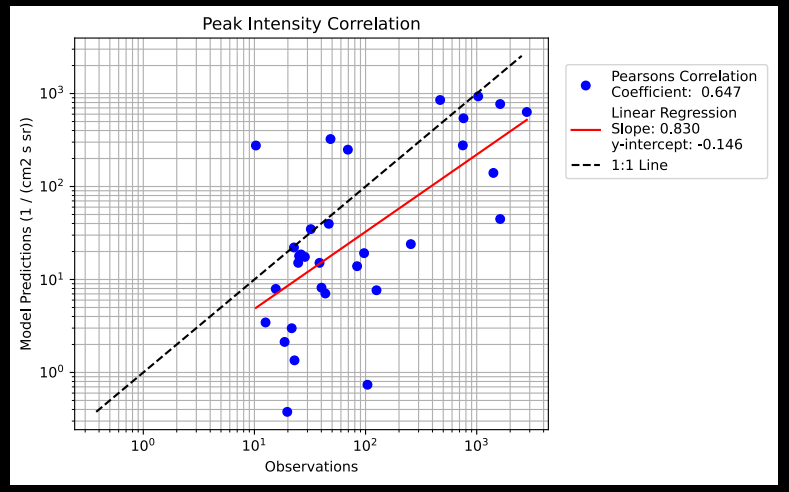
SEPVAL 2023: Onset Peak Summary >10 MeV, 10 pfu

- Comparison of higher and lower performing models for Onset Peak
  - High
  - Mid
  - Low

High



Mid



Low