Topical Discussion Meeting report

Name of the meeting: Flare Forecasting Research: Community Validation Tools
Conveners: KD Leka, Kathryn Whitman, Leila Mays (secretary)
Data – Time – Room: Date and time: Thursday 23nd November 2023 at 11:45-12:45 – Spot Room
Nr of participants: roughly 25 to 40

Objective of the TDM

Following the related CD100 session (Flare Forecasting Research: Where are we now?), flare forecasting is looking for a breakthrough. Community-supported tools may facilitate this research. We held a Topical Discussion Meeting for community engagement on the two following subjects: 1) Roles and Capabilities for a community validation and performance evaluation tool. What does this look like? What capabilities would it include? What options would be required, which would be nice to have? What input capabilities would be required? What products and output would be required or simply nice to have? What is the role of NASA/CCMC? What are some other appropriate (international) hosts? 2) Validation and Evaluation are only as good as the « answer » that is available. Every available flare event list has shortcomings. What should the community do to address this? Can we eliminate the repeated duplication of effort that seems to be happening? Can we design a curated, supported community-based solution with longevity ensured? How can we forward-think this for 4-Pi forecasting and validation? What (internationally-accepted) institution should or could host such a resource?

Some discussion highlights

- Discussed how a validation tool based on SPHINX-VIVID could be used for flare forecasting models.
- Looking at multiple metrics, providing thresholds and error bars for metrics, over long time periods, and different parts of the solar cycle. Some models perform well under some conditions but not others.
- Some models are predicting different time windows or have other differences and need to be carefully compared with others.
- There are many new flare forecasting models being built. How to evaluate them against existing models?
- Event list definitions are key to validation, but they have errors and everyone is independently correcting them.

Main conclusion of the meeting

- There is interest in providing a community validation tool as a web application for flare forecasting models. CCMC hosting the tool seems acceptable.
- It’s important for validation studies to show multiple metrics with error bars
- Community validation projects that compare models need to have a rules of the road and be built on trust
- There is an interest in having an international event list definition that is version controlled. Host is unclear, perhaps SIDC.

Annexes

Materials from discussion slides:

1) Roles and Capabilities for a community validation and performance evaluation tool.

a) **What does this look like?**

b) **What capabilities would it include?**

c) **What options would be required, which would be “nice to have”?**

d) **What input capabilities would be required?**

e) **What products and output would be required or simply nice to have?**

f) **What is the role of NASA/CCMC here? Are there other appropriate international hosts?**

2) Validation and Evaluation are only as good as the "answer" that is available.

a) **Every available flare event list has shortcomings.**

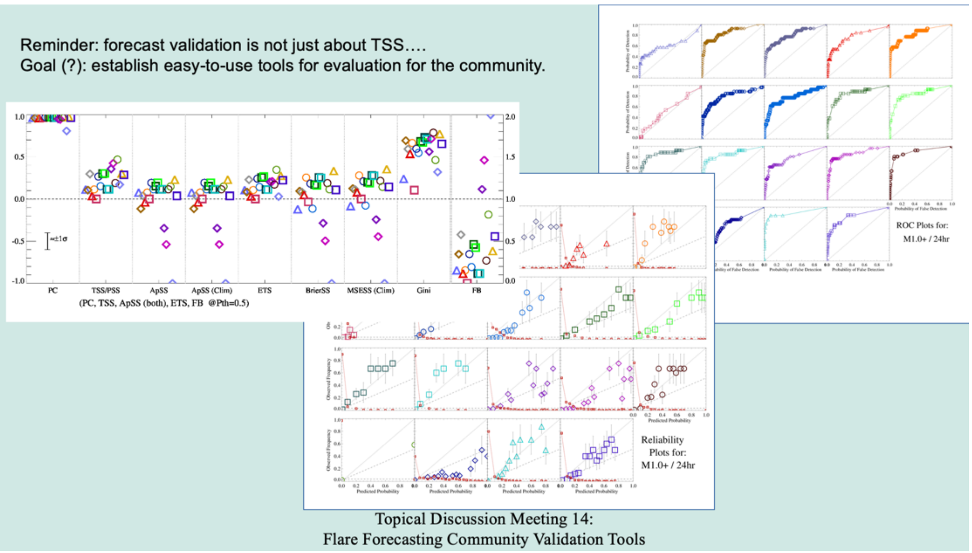
b) **What should the community do to address this?**

c) **Can we eliminate the repeated duplication of effort that seems to be happening?**

d) **Can we design a curated, supported community-based solution with longevity ensured?**

e) **How can we forward-think this for 4π forecasting and validation?**

f) **What (internationally-accepted) institution should or could host such a resource?**



Reminder: forecast validation is not just about TSS….
Goal (?): establish easy-to-use tools for evaluation for the community.

Topical Discussion Meeting 14:
Flare Forecasting Community Validation Tools

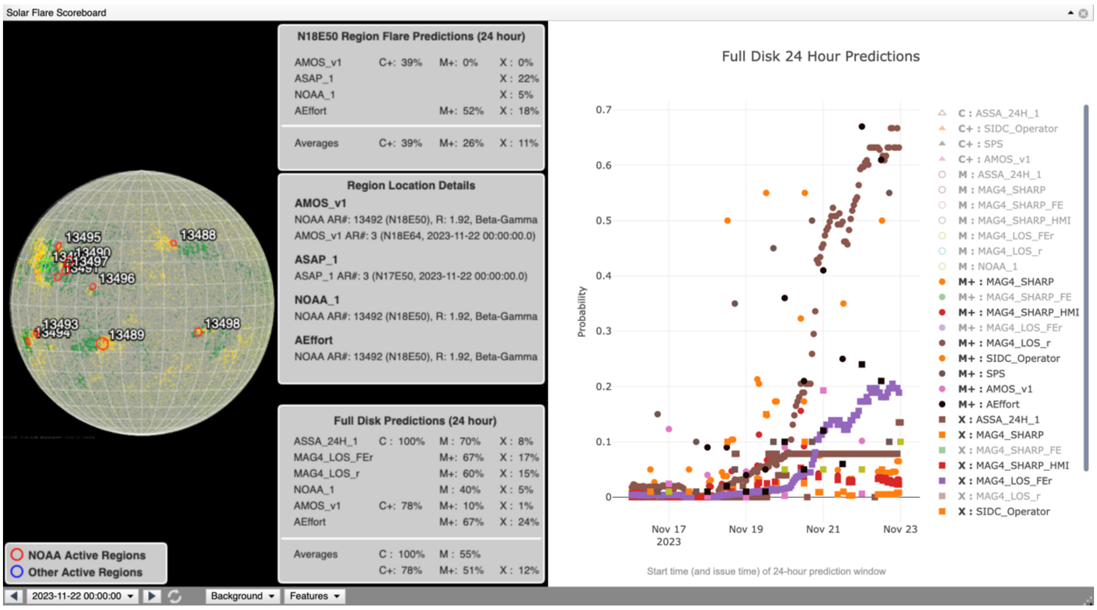
Community Forecasting Methods Scoreboards

Flare Scoreboard

SEP Scoreboard

CME Scoreboard
Submit your CME arrivable time predictions and compare with others.

- CCMC Scoreboards **collect and display** forecasts **before** the event is observed
- World-wide **ensemble** of research and operational models
- **Demonstrates** the operational potential of new capabilities
- Forecasts are freely **available** for community validation.
- Scoreboards do **not** focus on real-time validation, such as event and cumulative skill score calculations. These can be separate web applications, such as CAMEL and SPHINX-VIVID



Solar Flare Scoreboard

N18E50 Region Flare Predictions (24 hour)

AMOS_v1	C+: 39%	M+: 0%	X : 0%
ASAP_1			X : 22%
NOAA_1			X : 5%
AEffort		M+: 52%	X : 18%
Averages	C+: 39%	M+: 26%	X : 11%

Region Location Details

AMOS_v1
NOAA AR#: 13492 (N18E50), R: 1.92, Beta-Gamma
AMOS_v1 AR#: 3 (N18E64, 2023-11-22 00:00:00.0)

ASAP_1
ASAP_1 AR#: 3 (N17E50, 2023-11-22 00:00:00.0)

NOAA_1
NOAA AR#: 13492 (N18E50), R: 1.92, Beta-Gamma

AEffort
NOAA AR#: 13492 (N18E50), R: 1.92, Beta-Gamma

Full Disk Predictions (24 hour)

ASSA_24H_1	C : 100%	M : 70%	X : 8%
MAG4_LOS_FEr		M+: 67%	X : 17%
MAG4_LOS_r		M+: 60%	X : 15%
NOAA_1		M : 40%	X : 5%
AMOS_v1	C+: 78%	M+: 10%	X : 1%
AEffort		M+: 67%	X : 24%
Averages	C : 100%	M : 55%	
	C+: 78%	M+: 51%	X : 12%

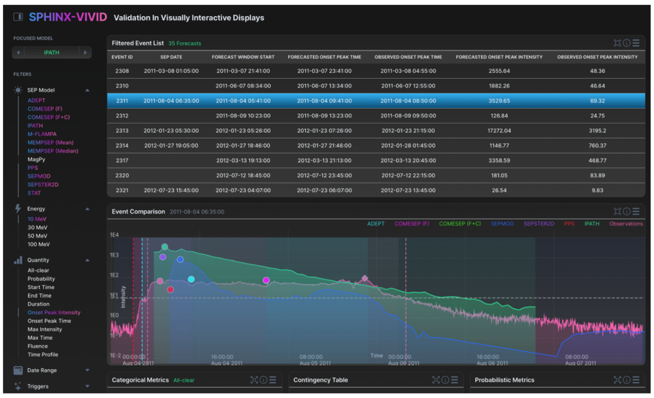
Validation In Visually Interactive Displays (VIVID)

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



SPHINX-VIVID Validation In Visually Interactive Displays

FOCUSED MODEL: IPATH

FILTERS

SEP Model: ADEPT, COMESEP (F), COMESEP (F+C), IPATH, M-FLAMPA, MEMPSEP (Mean), MEMPSEP (Median), MagPy, PPS, SEPMOD, SEPSTER2D, STAT

Energy: 10 MeV, 30 MeV, 50 MeV, 100 MeV

Quantity: All-clear, Probability, Start Time, End Time, Duration, Onset Peak Intensity, Onset Peak Time, Max Intensity, Max Time, Fluence, Time Profile

Date Range

Triggers

Filtered Event List 35 Forecasts

EVENT ID	SEP DATE	FORECAST WINDOW START	FORECASTED ONSET PEAK TIME	OBSERVED ONSET PEAK TIME	FORECASTED ONSET PEAK INTENSITY	OBSERVED ONSET PEAK INTENSITY
2308	2011-03-08 01:05:00	2011-03-07 21:41:00	2011-03-07 23:41:00	2011-03-08 04:55:00	2555.64	48.36
2310		2011-06-07 08:34:00	2011-06-07 13:34:00	2011-06-07 12:55:00	1882.26	46.64
2311	2011-08-04 06:35:00	2011-08-04 05:41:00	2011-08-04 09:41:00	2011-08-04 08:50:00	3529.65	69.32
2312		2011-08-09 10:23:00	2011-08-09 13:23:00	2011-08-09 09:50:00	126.84	24.75
2313	2012-01-23 05:30:00	2012-01-23 05:26:00	2012-01-23 07:26:00	2012-01-23 21:15:00	17272.04	3195.2
2314	2012-01-27 19:05:00	2012-01-27 18:46:00	2012-01-27 21:46:00	2012-01-28 01:45:00	1146.77	760.37
2317		2012-03-13 19:13:00	2012-03-13 21:13:00	2012-03-13 20:45:00	3358.59	468.77
2320		2012-07-12 18:45:00	2012-07-12 23:45:00	2012-07-12 22:15:00	181.05	83.89
2321	2012-07-23 15:45:00	2012-07-23 04:07:00	2012-07-23 06:07:00	2012-07-23 13:45:00	26.54	9.83

Event Comparison 2011-08-04 06:35:00

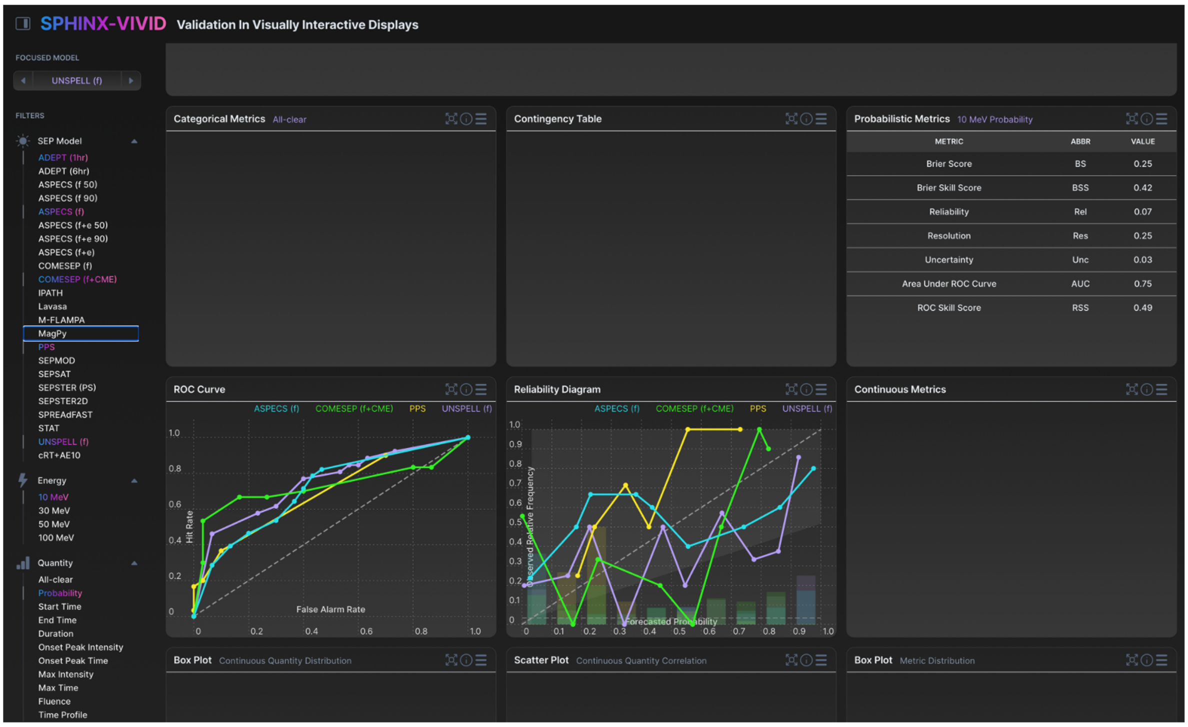
Categorical Metrics All-clear

Contingency Table

Probabilistic Metrics

Find poster during Thu/Fri poster session for more info

Visualization of Probability Metrics



SPHINX-VIVID Validation In Visually Interactive Displays

FOCUSED MODEL: UNSPELL (f)

FILTERS

SEP Model: ADEPT (1hr), ADEPT (6hr), ASPECS (f 50), ASPECS (f 90), ASPECS (f), ASPECS (f+e 50), ASPECS (f+e 90), ASPECS (f+e), COMESEP (f), COMESEP (f+CME), IPATH, Lavasa, M-FLAMPA, MagPy, PPS, SEPMOD, SEPSAT, SEPSTER (PS), SEPSTER2D, SPREAdFAST, STAT, UNSPELL (f), cRT+AE10

Energy: 10 MeV, 30 MeV, 50 MeV, 100 MeV

Quantity: All-clear, Probability, Start Time, End Time, Duration, Onset Peak Intensity, Onset Peak Time, Max Intensity, Max Time, Fluence, Time Profile

Categorical Metrics All-clear

Contingency Table

Probabilistic Metrics 10 MeV Probability

METRIC	ABBR	VALUE
Brier Score	BS	0.25
Brier Skill Score	BSS	0.42
Reliability	Rel	0.07
Resolution	Res	0.25
Uncertainty	Unc	0.03
Area Under ROC Curve	AUC	0.75
ROC Skill Score	RSS	0.49

ROC Curve

Reliability Diagram

Continuous Metrics

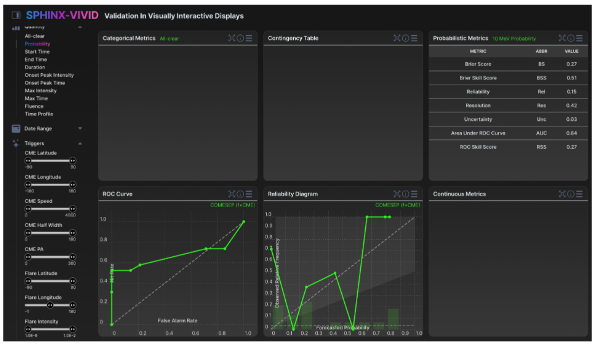
Box Plot Continuous Quantity Distribution

Scatter Plot Continuous Quantity Correlation

Box Plot Metric Distribution

On-the-Fly Filtering

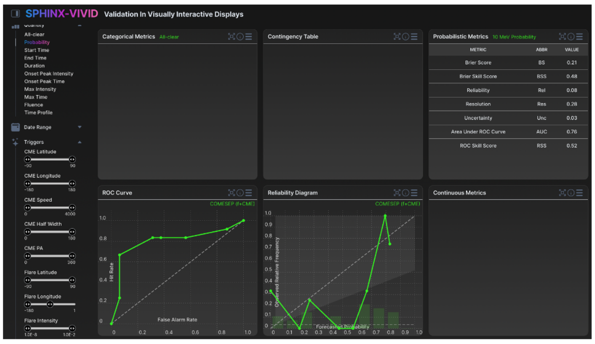
Flares from Eastern Longitudes (< -1 degrees)



Probabilistic Metrics 10 MeV Probability

METRIC	ABBR	VALUE
Brier Score	BS	0.27
Brier Skill Score	BSS	0.51
Reliability	Rel	0.15
Resolution	Res	0.42
Uncertainty	Unc	0.03
Area Under ROC Curve	AUC	0.64
ROC Skill Score	RSS	0.27

Flares from Western Longitudes (> -1 degrees)



Probabilistic Metrics 10 MeV Probability

METRIC	ABBR	VALUE
Brier Score	BS	0.21
Brier Skill Score	BSS	0.48
Reliability	Rel	0.08
Resolution	Res	0.28
Uncertainty	Unc	0.03
Area Under ROC Curve	AUC	0.76
ROC Skill Score	RSS	0.52

Note that only 32 SEP events and 30 non-event periods are included in this analysis → Low statistics

Discussion contributors:

KD Leka
Katie Whitman
Phil Quinn
Leila Mays
Manolis Georgolis
Robert Jarolim

Paul Wright
Chris Light
Karin Dissauer
Larisza Krista
Kim Moreland

Minutes:

KD: Introduced the TDM discussion topics
Leila: Introduced CCMC Scoreboards, Flare Scoreboard, planned GUI upgrades, and need for a separate flare-forecast validation tool.
Katie Whitman and Phil Quinn: Introduced SPHINX-VIVID, a web application to validate and visualize SEP models (this project is related to SEPVAL meetings and the ISWAT team).

Discussion:

Manolis: What period and how often is data loaded into SPHINX. This is a great tool for seeing model performance dynamically for different periods of the solar cycle (SC) and other ways of splitting up the time period.
Phil: Here, a specific SEPVAL study time period is loaded, could load a different study.
Leila: For example we could load up the Nagoya flare workshop data [ed: Leka+2019] and then check a new model or upgrade against these.
KD: A different time period would be better for looking at scores because of it being a short time period [ed: and also solely full-disk forecasts with limited options for event definitions], but the methodology holds.

KD: Posed the question – how would people like to interact with such a validation tool? Should there be an established onboarding process to get access to use the tool? Or would it be advantageous to

have a “quick look” or private-use area for a user to upload (e.g.) their probability results and relevant metadata to explore and/or compare with other forecast results that are available publicly? What about the event list to be used by such a tool?
Katie: There is a need for a curated event list that is continuously updated with corrections and new events.
KD: This is the essence of Question #2 - How to get international buy-in for an event list so that it is accepted (trusted) by everyone that it is complete, and not biased?

Robert: How would such a validation tool deal with models having different objectives, such as a shorter (6 hr) forecast prediction window (validity period) vs 24 hr. Should these be split up? They can’t really be compared.
Phil: This can be built into the tool, so that you can only select a single prediction window definition at a time when computing scores/plotting.
Leila: Presently, the CCMC Flare scoreboard prediction window is 24 hrs; if a model has another window we don’t plot it with the others because it doesn’t have the same meaning [ed: different event definition] which can be confusing to users [ed: not valid to directly compare]. For example, ASSA model has a 12 hr prediction window which was converted to 24 hr (method provided by model developer) in order to display it on the scoreboard.

Manolis: There is a relation between the work done by the flare scoreboard with the format and display, and a validation tool. Scoreboard is collecting the data which can then be used in VIVID. Let’s take advantage of the work already done here and build on it.

KD: At what point should a new method/upgrade be onboarded into a tool like this? Should it be private? Researchers could use this to explore performance, and the interplay of metrics. Or will this lead to some people misrepresenting their model as performing better than it actually is?
Paul: Having participated in ML modelling groups where score results are submitted, there are situations where models are run & run to “get to perfect”, everyone is just trying to be at the top to look better, but then it’s unclear their actual performance. This is an issue.
KD: Nagoya workshop had a rules of the road and multiple metrics and context was provided, important to have something like this.
Chris: If exploration of their model results are private and they are optimizing behind the scenes, will people really be fooled? Can’t they see through this?
Paul: Yes, but there isn’t enough time to read all of the papers, and the information is hard to find. Perhaps, for question (1a) even just collecting all the flare forecasting papers in one place would be useful.
Manolis: This is an issue, right now the journals are flooded with flare forecast models and it’s not clear their value. Also how good is good, how bad is bad? What are the most important metrics for flares, CMEs and SEPs?

KD: The metrics are on probabilities, and usually the TSS is reported, with no threshold provided. Can fool people but it would be hard to full people across all the metrics a system like this calculates, so this could help with the issue.

Karin: Should we require models to go through this tool before being onboarded on the scoreboard.
Leila: That’s possible, or perhaps instead of a gatekeeper, this is a tool that allows them to do the analysis (they may not have explored with their own tools) and/or comparison with other studies. Or they could download the data from those studies and use their own validation tools. Currently, any model can be added to the [CCMC] flare scoreboard, there is no performance requirement. In fact it helps to bring them in and really see what their performance is like in real-time compared to the scores they report.

Manolis: from the Nagoya flare workshop, I learned about the Appleman Skill Score and what it means. Multiple scores are needed. So this kind of tool and collaboration with others could be helpful.
Chris: May not require the community to use the tool, but could require them to report multiple scores instead of just one.
Larisza: The proof in the pudding is eating it. From the community coronal hole (CH) validation they saw that some CH models had high scores and good performance for some time periods, but when they tested other periods in different parts of the SC, performance was less good.
For flare validation, we need long continuous time periods throughout SCs and different data sources. Large sample size

KD: Large sample size is important, smaller is not invalid, but error bars on the scores should reflect the sample size. This is needed for VIVID [ed: this improvement is planned by Phil].
Karin: is there a SC database we could use to define the different parts of the SC? Use this together with an event list that needs to span SCs.

KD: Should a validation tool like VIVID be hosted by CCMC, or is there a concern that this is too NASA or CCMC focused, would people be comfortable with this?
Katie: Or for example, should we develop the tool in consultation with ESA and other groups to get their buy in?
Paul: As long as there is a good API, personally ok with CCMC hosting.
However, if NASA, CCMC is ranking and comparing the models, will that be ok politically?
Leila: Trust with modelers is important. There needs to be a rules of the road for ranking and comparing.
Paul: trust is important, however if a model is deemed “bad” that could impact someone’s career built around this model.
Leila: Yes this is a concern that needs to be accounted for. CCMC has been doing challenges for years and this was a worry at first. We built trust and kept results to the validation group until verified to be correct and caveated in the report and received permission from the developer. [ed: question to follow up re: what kind of “report” would be expected / provided, required to be public at which stage, or effectively the “permanence” of running the validation tool?]
Larisza: Also a model may be “bad” in the period under study but not bad in another circumstance, this needs to be clearly explained.

KD: Discussion of item 2 – all validation depends on the event list, the “ground truth”. Some teams only want to validate against their preferred list. Teams around the world are taking an event list and then correcting it individually, everyone is duplicating efforts. The corrections are sometimes not available back to the community. How could we work together to curate an international event list?
Manolis: We need another workshop to continue this work. Make event lists for each problem we want to study, and define the metrics for each problem. Come to these via community consensus.
KD: Agree and should be done, but this is ambitious! As a first step, let’s start with a version controlled flare event list. Where should such a list be hosted? Needs to be curated and allow new submissions/edits from anyone.
Larisza: NOAA lists do need cleaning, but it should be NOAA that provides the gold standard, because the flare definitions are from GOES [ed: a NOAA asset]. Also a correction is coming - GOES magnitudes will be rescaled. They want to hear about the corrections.
KD: Unfortunately reported corrections are not making it back to the NOAA list. This problem needs to be addressed.

Kim: How about starting by comparing the existing flare lists?
KD: Yes – also SIDC has a relational database infrastructure where edits are version controlled and

timestamped. Suggest starting with existing lists, maybe 80% of the database will be done. The community would edit and clean the remainder. Any user should be able to request/suggest an edit. How to fund a core group of curators to check edits and new additions? Who should host this? Important to also track which of these event list versions are used in the validation tool.
Larisza: Needs to be a conversation with NOAA. NOAA is also making corrections to the list both into the future but backwards.
KD: NOAA is a good starting location but more is needed, like the corrections and timestamping/versioning. Need a way to see when/what NOAA updates in their list. The SIDC relational database entries are timestamped [ed: MetOffice, SolarMonitor, other institutions may have been tracking updates too.]
Paul: Oftentimes the result of updates and corrections (NOAA and Stanford merge for example) is buried in a paper. Need to gather all of these.
KD: May also need to validate against different event list definitions, such as definitive and near-real-time.
Manolis: Validating against definitive data should be good enough to get an idea of how it woud perform against a real-time event definition.

KD/Leila/Katie: thank you all for your ideas, your thoughts, your insights. Please feel free to contact any of us with further thoughts, and we will provide updates to the community as things develop.