

# Operational Flare Forecasting: Benchmarks and Initial Performance Comparisons



Workshop held late 2017 at Nagoya University  
hosted by the  
Institute for Space-Earth Environmental Research (ISEE)'s  
Center for International Collaborative Research (CICR).



## Methodology and Initial Results

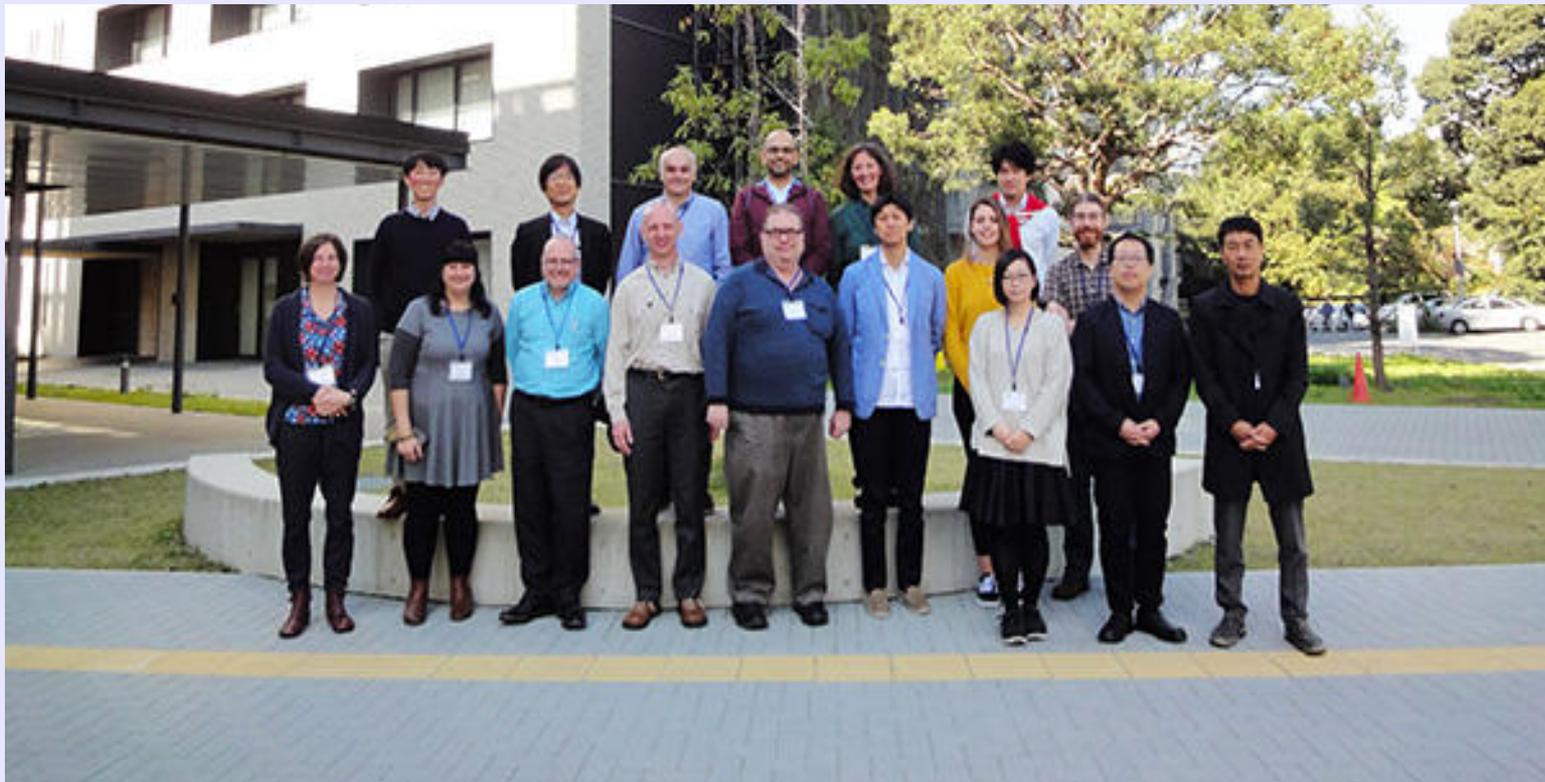
(slides prepared by KD Leka for the International Flare Forecasting  
Comparison 3 (IFFC3) Team)

### GOALS

- Quantitatively evaluate the performance of today's methods, both established and new.
- Establish benchmarks for evaluating future methods.
- Investigate what approaches perform better to enable future improvements

# MOTIVATION AND BACKGROUND

- How well do operational Flare Forecasting methods presently work?
- New methods are being developed and tested on the most recent (albeit fairly quiet) cycle, and a few research methods have been brought to full operational capability. Do they improve upon the established forecasts?
- What is needed to quantitatively answer that question to begin with?



*"unencumbered science"*

## METHODS & TEAM

The forecasting facilities which were selected and invited were *operational*, meaning that they “provide forecasts on a routine, consistent basis using only data available prior to the issuance time.” Human intervention did not disqualify, but daily forecasts were expected and methods were penalized when they were not available.

NOAA/SWPC	Rob Steenburgh
MetOffice MOSWOC	Suzy Bingham, Mike Sharpe
NICT	Yuki Kubo
MAG4 [LOS/Vect, W and WF]	David Falconer
ASAP	Tarek A.M.Hamad Nageem, Rami Qahwaji
ASSA	JunChul Mun, Sangwoo Lee
NJIT	Ju Jing, Sung-Hong Park
A-EFFORT	Manolis Georgoulis
BoM/SWS (FlareCast and Climatology)	Mike Terkildsen, Graham Steward
SIDC (Royal Obs. Belgium)	Jesse Andries, Veronique Delouille
AMOS	Kangjin Lee
DAFFS, DAFFS-G	Graham Barnes, KD Leka
MCSTAT, MCEVOL	Shaun Bloomfield, Aoife McCloskey, Sophie Murray, Peter Gallagher
(consulting)	Yumi Bamba, M. Leila Mays, Kanya Kusano

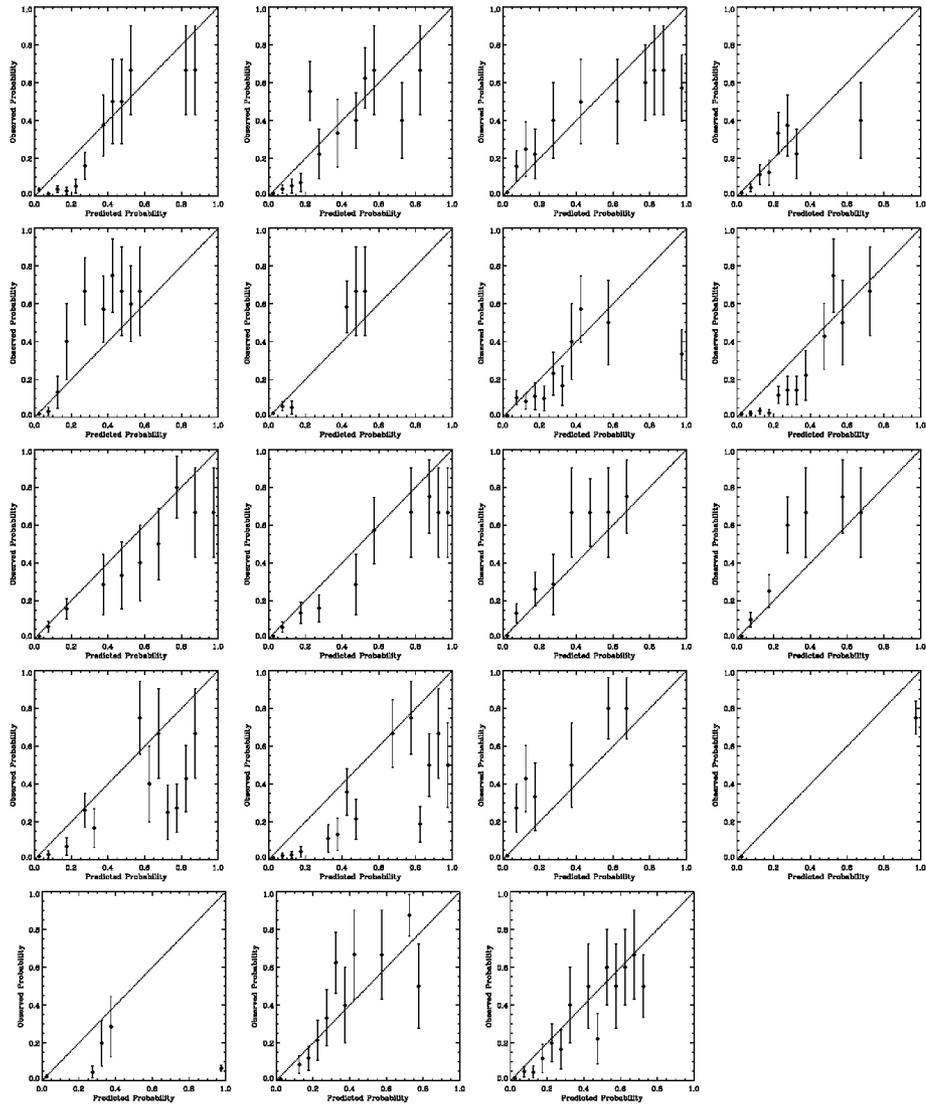
# METHODOLOGY for fair, head-to-head benchmarking:

- *Establish an agreed-upon forecast test interval:*
  - 2016 – 2017 inclusive
  - Long enough (????) for sample sizes, allows sufficient (????) training interval for methods using solely SDO-era data.
- *Full-disk forecasts only*
  - Mitigates issues of differing region definitions
  - Only one method does *not* produce full-disk forecasts, and the component forecasts were combined post-facto into full-disk forecasts
- *Agreed-upon Event Definitions*
  - Lower-limits plus Exceedance:
    - C1.0+, M1.0+, X1.0+
    - Those methods not producing exceedance forecasts were converted to exceedance forecasts using conditional probabilities over a method's training interval.
    - Not all methods produce smaller-event forecasts.
  - 24-hr validity period
    - Most methods provide a forecast at or near 00:00 UT
    - No attempt to correct hour-or-so discrepancies
  - 0hr latencies
    - Many methods produced 24hr, 48hr latency forecasts as well; to be evaluated later.
- **RESULTING SAMPLE SIZES ARE NOT GOOD:**
  - M1.0+ : 26 events, 705 non-events.
    - Not every method did C1.0 or C1.0+ forecasts
    - Not even bothering with X1.0+ (3 events in testing interval).

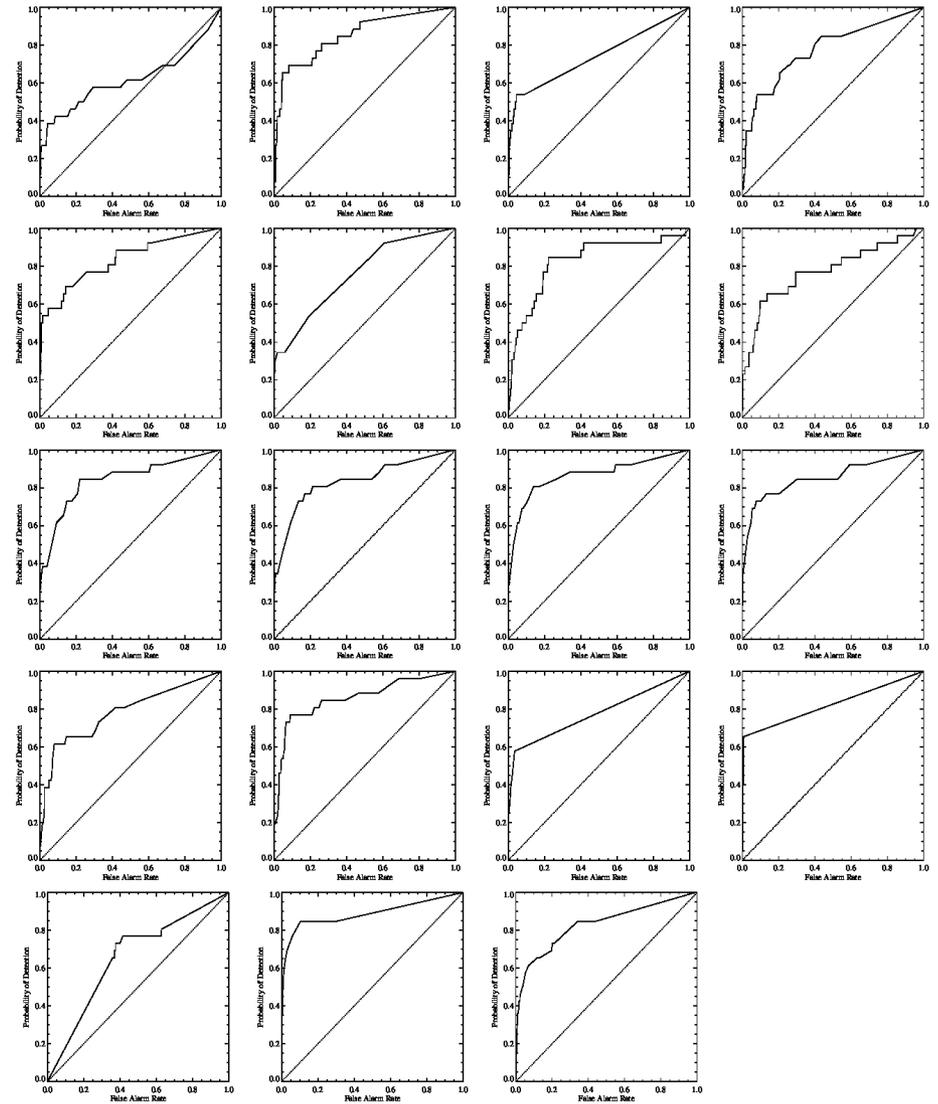
## METRICS and NOTES:

- *A variety of evaluation plots and metrics used*
  - Focus on metrics that do not depend on categorical forecasts:
    - Brier Skill Score and Gini Coefficients, to summarize Reliability plots and Receiver Operating Characteristic (ROC) plots, respectively.
    - All but one method produces probabilistic forecasts
      - Categorical forecast → probability by assigning [0%, 100%] probabilities.
  - Include categorical-based metrics for now (likely not for publication)
    - TrueSkillStatistic, Appleman, RateCorrect and Heidke
    - A “Probability Threshold” (above which a forecast is made for an event to occur) is required for categorical-based metrics.
      - $P_{\text{thresh}}=0.5$  chosen by default.
- Missing forecasts were assigned Probability=0.0.
  - Operational forecasting paradigm.

# Reliability Plots



# ROC curves

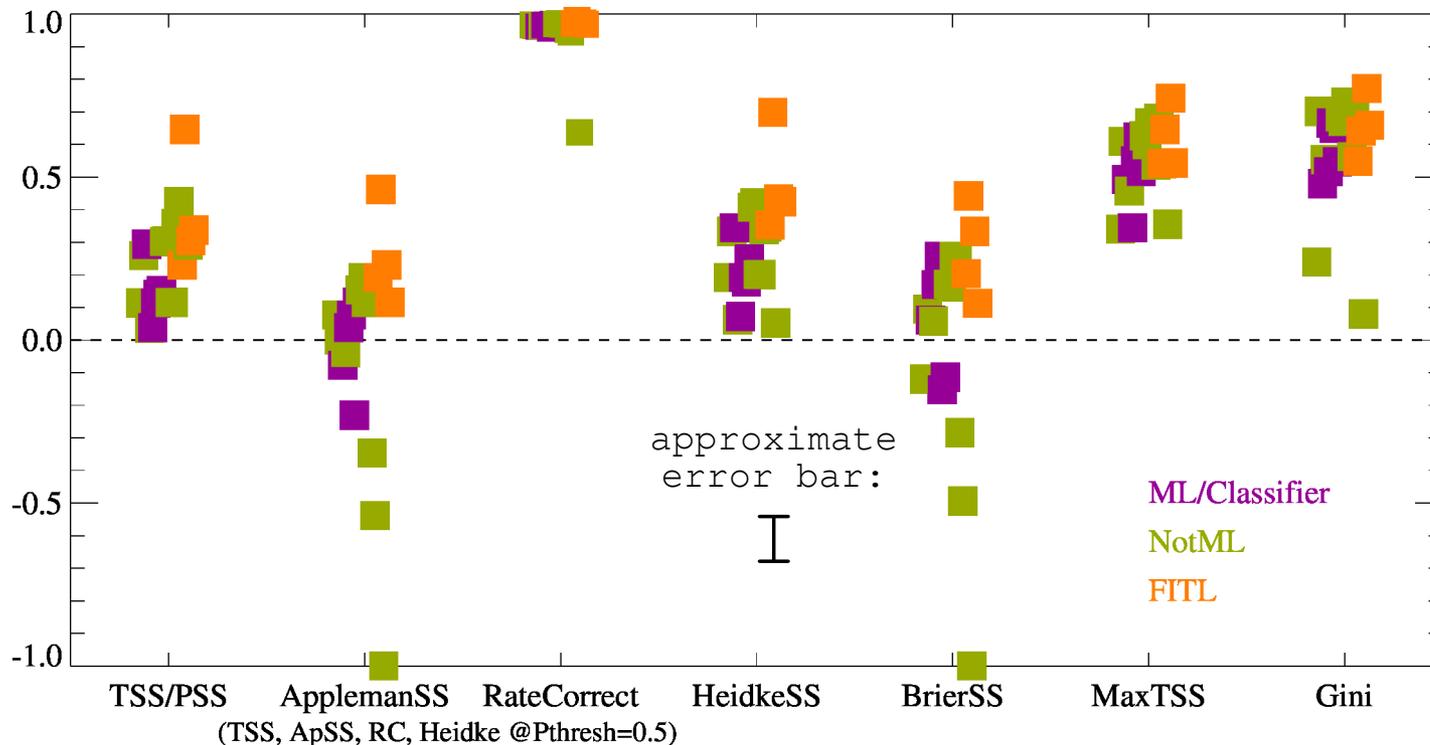


*None look particularly wonderful....*

# EXAMPLE ANALYSIS: Forecast Production

Describes algorithm used to produce the forecast, whether a statistical classifier (including machine learning), a statistical non-classifier (Poisson statistics, correlation curves), or includes a human (even if other approaches are used as well).

Machine Learning / Classifier	DAFFS & DAFFS-G, ASAP, BoM/SWS
Not Machine Learning	MAG4, A-EFFORT, NJIT, ASSA, MCSTAT, MCEVOL, NOAA, MetOffice, AMOS
Forecaster-in-the-Loop	MOSWOC, NOAA, SIDC, NICT



Various metrics of the methods coded by type of forecast production.

Some indication that forecaster-in-the-loop is advantageous.

Not definitively separated between classifications.