The Astrophysical Journal Supplement Series, 243:36 (15pp), 2019 August https://doi.org/10.3847/1538-4365/ab2e12

A Comparison of Flare Forecasting Methods. II. Benchmarks, Metrics, and Performance Results for Operational Solar Flare Forecasting Systems

K. D. Leka[1,2], Sung-Hong Park[1], Kanya Kusano[1], Jesse Andries[3], Graham Barnes[2], Suzy Bingham[4], D. Shaun Bloomfield[5], Aoife E. McCloskey[6], Veronique Delouille[3], David Falconer[7], Peter T. Gallagher[8], Manolis K. Georgoulis[9,10], Yuki Kubo[11], Kangjin Lee[12,13], Sangwoo Lee[14], Vasily Lobzin[15], JunChul Mun[16], Sophie A. Murray[6,8], Tarek A. M. Hamad Nageem[17], Rami Qahwaji[17], Michael Sharpe[4], Robert A. Steenburgh[18], Graham Steward[15], and Michael Terkildsen[15]

[1] Institute for Space-Earth Environmental Research, Nagoya University, Furo-cho Chikusa-ku Nagoya, Aichi 464-8601, Japan; kdleka@isee.nagoya-u.ac.jp, leka@nwra.com
[2] NorthWest Research Associates, 3380 Mitchell Lane, Boulder, CO 80301, USA
[3] STCE—Royal Observatory of Belgium, Avenue Circulaire, 3 B-1180 Brussels, Belgium
[4] Met Office, FitzRoy Road, Exeter, Devon, EX1 3PB, UK
[5] Northumbria University, Newcastle upon Tyne, NE1 8ST, UK
[6] School of Physics, Trinity College Dublin, College Green, Dublin 2, Ireland
[7] NASA/NSSTC, Mail Code ST13, 320 Sparkman Drive, Huntsville, AL 35805, USA
[8] School of Cosmic Physics, Dublin Institute for Advanced Studies, 31 Fitzwilliam Place, Dublin, D02 XF86, Ireland
[9] Department of Physics & Astronomy, Georgia State University, 1 Park Place, Rm #715, Atlanta, GA 30303, USA
[10] Academy of Athens, 4 Soranou Efesiou Street, 11527 Athens, Greece
[11] National Institute of Information and Communications Technology, 4-2-1 Nukuikita Koganei, Tokyo 184-8795, Japan
[12] Meteorological Satellite Ground Segment Development Center Electronics and Telecommunications Research Institute, Daejeon 218 Gajeong-ro, Yuseong-gu, Daejeon, 34129, Republic of Korea
[13] Kyung Hee University, 1732 Deogyeong-daero, Giheung-gu, Yongin, 17104, Republic of Korea
[14] SELab, Inc., 150-8, Nonhyeon-ro, Gangnam-gu, Seoul, 06049, Republic of Korea
[15] Bureau of Meteorology, Space Weather Services, P.O. Box 1386, Haymarket NSW 1240, Australia
[16] Korean Space Weather Center, 198-6, Gwideok-ro, Hallim-eup, Jeju-si, 63025, Republic of Korea
[17] University of Bradford, Bradford West Yorkshire BD7 1DP, UK
[18] NOAA/National Weather Service National Centers for Environmental Prediction Space Weather Prediction Center, W/NP9, 325 Broadway, Boulder, CO 80305, USA

Received 2019 March 28; revised 2019 June 18; accepted 2019 July 1; published 2019 August 16

Abstract

Solar flares are extremely energetic phenomena in our solar system. Their impulsive and often drastic radiative increases, particularly at short wavelengths, bring immediate impacts that motivate solar physics and space weather research to understand solar flares to the point of being able to forecast them. As data and algorithms improve dramatically, questions must be asked concerning how well the forecasting performs; crucially, we must ask how to rigorously measure performance in order to critically gauge any improvements. Building upon earlier-developed methodology of Paper I (Barnes et al. 2016), international representatives of regional warning centers and research facilities assembled in 2017 at the Institute for Space-Earth Environmental Research, Nagoya University, Japan to, for the first time, directly compare the performance of operational solar flare forecasting methods. Multiple quantitative evaluation metrics are employed, with the focus and discussion on evaluation methodologies given the restrictions of operational forecasting. Numerous methods performed consistently above the "no-skill" level, although which method scored top marks is decisively a function of flare event definition and the metric used; there was no single winner. Following in this paper series, we ask why the performances differ by examining implementation details (Leka et al. 2019), and then we present a novel analysis method to evaluate temporal patterns of forecasting errors in Paper IV (Park et al. 2019). With these works, this team presents a well-defined and robust methodology for evaluating solar flare forecasting methods in both research and operational frameworks and today's performance benchmarks against which improvements and new methods may be compared.

Key words: methods: data analysis – methods: statistical – Sun: activity – Sun: flares – Sun: magnetic fields

1. Introduction

Solar flares can be considered the initiating event for many space weather phenomena and impacts. The impact of solar flare radiation is almost immediate in the case of sudden ionospheric disturbances, particularly with M- and X-class flares, which disrupt radar and terrestrial communications systems in the sunlit hemisphere. Solar flares are also intimately associated with other pertinent space weather phenomena, such as energetic particle storms and coronal mass ejections whose impacts may be delayed relative to flare impacts, but can incur broader effects. Predicting solar flare likelihood has thus long been a defined and required operational product, now with several facilities worldwide providing operational forecasts to a variety of customers.

Predicting solar flares is also the ultimate test of understanding their cause or causes. They have long been associated with certain morphological aspects of solar active regions, such as complex structures, strong-gradient polarity inversion lines, and indications of significant energy storage in the magnetic field itself (see, e.g., Sawyer et al. 1986 and references herein; Leka & Barnes 2003; Schrijver 2007). The only appropriate energy source is the stored free magnetic energy in solar active

region magnetic fields, and the only appropriate release mechanism invokes magnetic reconnection and reconfiguration to release that free magnetic energy. Indeed, as discussed below and further in Leka et al. (2019, hereafter Paper III), quantitative "modern" forecasts incorporate this physical understanding as they often characterize coronal magnetic energy storage by proxy, with the parametrizations of photospheric magnetograms. In these contexts, however, pinpointing a unique triggering mechanism has remained elusive. Alternatively, solar flares may inherently be stochastic in nature (see, e.g., Wheatland 2000; Strugarek et al. 2014; Aschwanden et al. 2016) thus essentially unpredictable in a deterministic sense. The state of the research is presently at a point where it is still unknown in which regime the physics operates. While their heliospheric and societal impacts provide motivation for predicting these energetic events, success or failure at forecasting also provides a key indicator as to whether stochastic physics is or is not involved.

In 2009, the first in a series of workshops was held to compare and evaluate the newly emerging plethora of methods aimed at distinguishing those solar active regions with an imminent flare threat. Data from the *Solar and Heliospheric Observatory* (*SoHO*; Domingo et al. 1995) and specifically the Michelson Doppler Imager (MDI; Scherrer et al. 1995) were provided to the methods for an analysis. The performance results (see Barnes et al. 2016, hereafter Paper I) are of secondary importance to the methodology that was established, identifying the importance of common definitions and standard metrics when determining what constitutes "good performance."

During Solar Cycle 24, the availability of significantly improved data sources, such as the Helioseismic and Magnetic Imager (HMI) on the *Solar Dynamics Observatory* (*SDO* Pesnell et al. 2012; Scherrer et al. 2012; Schou et al. 2012; Centeno et al. 2014; Hoeksema et al. 2014) has made possible a growing variety of flare forecasting systems that are running in an operational mode (some of which were in the development phase in 2009). Consequently, an international collaboration effort was initiated through the Center for International Collaborative Research (CICR), at the Institute for Space-Earth Environmental Research (ISEE), Nagoya University, Japan, to bring together the operational forecasting teams from a variety of institutions (government, private, and academic) to evaluate the performance of different techniques. The goals of that workshop and the subsequent analysis are to (1) establish benchmarks and comparison methodologies for operational flare-forecasting facilities and (2) begin to understand what particular forecasting methodologies enable the best forecasting performance.

The participating systems are listed in Section 2 with additional relevant (unpublished) details elaborated upon in Appendix A. Although additional research into improving forecasts is being published frequently as of late (Bobra & Couvidat 2015; Nishizuka et al. 2017; Florios et al. 2018), for this research the comparisons were limited to those truly running in an operational manner, which the group describes as providing a forecast on a routine, consistent basis using only data available prior to the issuance time. Many methods, especially the long-standing governmental–institutional methods, rely on sunspot classification and historical flaring rates (Sawyer et al. 1986; McIntosh 1990). A few are now employing more sophisticated analyses of the host sunspot groups and statistical classifiers or machine-learning algorithms. Forecasts were not required to be fully automatic; human intervention, i.e., a "forecaster in the loop" (FITL), was explicitly allowed. Providing a forecast on a daily basis was also not a requirement, although as an operational system, not doing so was effectively penalized by the evaluation metrics, as described in Section 2.2. No further restrictions were placed on the data employed or interval used for training, except that it could not overlap with the testing interval (see Section 2.1). The impacts of long- versus short-training intervals (e.g., whether more than one solar cycle was used for training the method) and other details are discussed further in Paper III.

The participants provided forecasts for an agreed-upon interval with agreed-upon event definitions as described in Section 2.1 (Leka & Park 2019). Representatives from most of the participating groups attended (in person or remotely) a three-day workshop during which the approaches and initial results were discussed in depth. The results of those days, plus further discussions and analyses that occurred in the subsequent months, are now presented here and in Paper III and Park et al. (2019, hereafter Paper IV).

2. Comparison Methodology

The participating facilities and methods (with their monikers and published references, as available) are listed in Figure 1, and specific details that are not available from published literature (or modifications that have been made since the relevant publications) are briefly described in Appendix A. Some methods have multiple options for producing forecasts, and those are also delineated both in Figure 1 and Appendix A. In Paper III, we distinguish the methods according to broad categorizations of their implementations, such as data sources, training intervals, imposed limits, forecast approach (e.g., statistical, FITL), etc., and hence we leave that level of detail to that paper.

2.1. Event Definitions and Testing Interval

The participants agreed on a testing interval of 2016 January 1–2017 December 31 for evaluating forecasts. This is arguably a very short testing interval; in the present situation, it was chosen to balance both training and testing data for those methods relying on data from *SDO*/HMI, since the near real time (NRT) data from HMI are only available from late 2012. The resulting activity levels are summarized in Table 1. Evaluation was performed on full-disk forecasts only to avoid the requirement of standardizing the different active region identification methods in use (combining region-based forecasts to the full disk is described in Appendix B.1).

Event definition choices were dictated by the need for common definitions across methods and the fact that these are operational methods, hence most already produce forecasts that match the NOAA/Space Weather Prediction Center (SWPC)-established event definition and timings.

As such, the group agreed upon event thresholds as "lower-limits plus exceedance" following the NOAA/SWPC definition, based on the NOAA *Geostationary Observing Earth Satellite* (*GOES*) X-Ray Sensor (XRS) 1–8 Å bands, `C1.0+` and `M1.0+`, corresponding to lower limits of 1.0×10^{-6} and 1.0×10^{-5} Wm^{-2}, respectively, with no upper limit (i.e., "exceedance" forecasts). All forecasts were put onto an exceedance basis; calculating exceedance forecasts from

Institution	Name of Method/Code[a]	Label	Symbol	Reference(s)
ESA/SSA A-EFFORT Service	Athens Effective Solar Flare Forecasting	A-EFFORT		Georgoulis & Rust (2007)
Korean Meteorological Administration & Kyung Hee University	Automatic McIntosh-based Occurrence probability of Solar activity	AMOS		Lee et al. (2012)
University of Bradford (UK)	Automated Solar Activity Prediction	ASAP		Colak & Qahwaji (2008, 2009)
Korean Space Weather Center (by SELab, Inc)	Automatic Solar Synoptic Analyzer	ASSA		Hong et al. (2014), Lee et al. (2013)
Bureau of Meteorology (Australia)	Flarecast II	BOM		Steward et al. (2011, 2017)
120-day No-skill Forecast	Constructed from NOAA event lists	CLIM120		Sharpe & Murray (2017)
NorthWest Research Associates (USA)	Discriminant Analysis Flare Forecasting System	DAFFS		Leka et al. (2018)
" "	GONG+*GOES* only	DAFFS-G		" "
NASA/Marshall Space Flight Center (USA)	MAG4 (+according to	MAG4W		Falconer et al. (2011);
" "	magnetogram source	MAG4WF		also see Appendix A
" "	and flare-history	MAG4VW		
" "	inclusion)	MAG4VWF		
Trinity College Dublin (Ireland)	SolarMonitor.org Flare Prediction System (FPS)	MCSTAT		Gallagher et al. (2002), Bloomfield et al. (2012)
" "	FPS with evolutionary history	MCEVOL		McCloskey et al. (2018)
Met Office (UK)	Met Office Space Weather Operational Center human-edited forecasts	MOSWOC		Murray et al. (2017)
National Institute of Information and Communications Technology (Japan)	NICT-human	NICT		Kubo et al. (2017)
New Jersey Institute of Technology (UK)	NJIT-helicity	NJIT		Park et al. (2010)
NOAA/Space Weather Prediction Center (USA)		NOAA		Crown (2012)
Royal Observatory Belgium Regional Warning Center	Solar Influences Data Analysis Center human-generated	SIDC		Berghmans et al. (2005), Devos et al. (2014)

a: if applicable

Figure 1. Participating Operational Forecasting Methods (Alphabetical by Label Used).

Table 1
24 hr Event Rates for 2016 January 1–2017 December 31

Class	# of Quiet Days	# of Event Days	Climatology (Event Day Rate)
C1.0+	543	188	0.257
M1.0+	705	26	0.036
X1.0+	728	3	0.004

category-limited forecasts (i.e., including an upper limit), as were provided by some methods, is discussed in Appendix B.2. No background or pre-flare subtraction was performed for the evaluation data, which is consistent with none generally being performed by any operational method during either the training or event prediction (see also Wheatland 2005 for a discussion on the impact of background subtraction). The event definitions include 24 hr validity periods and effectively 0 hr latencies (the time periods between forecast issuance and the start of the validity period) for the initial comparisons (i.e., only one-day forecasts, not longer-range forecasts). Longer effective latencies may be implied due to data acquisition times, but these are ignored here for delays <1 hr. Additionally, note that a number of centers produce additional forecasts (with variations in frequency of forecast, event thresholds, latencies, or validity periods); for this comparison, we chose the event definitions to assure the most overlap between methods. We refer now to these two event definitions using the shorthand "C1.0+/0/24" and "M1.0+/0/24," noting that the nomenclature includes all three aspects of the event definition (thresholds, latency in hours, and validity period in hours).

The C1.0+/0/24 exceedance definition provided 188 event days, and the M1.0+/0/24 exceedance definition provided 26 event days over the 731 days of the testing interval (2016 was a leap year; see Table 1). Not all methods produce C1.0+/0/24 forecasts. While most methods produce a forecast for X1.0+ (1.0×10^{-4} Wm^{-2} and larger), in practice, the short testing interval produced too few of these largest events to provide meaningful evaluations.

Most methods issue a forecast in the neighborhood of 00:00 UT. Within approximately one hour, any discrepancy from midnight was ignored. Beyond that, the discrepancies in event lists would become problematic. For methods which issue forecasts significantly different from midnight (SIDC at 12:30 UT, NICT at 06:00 UT), custom event lists were constructed based on that issuance time. Although these custom lists do change the number of events slightly (C1.0+/0/24 becomes 183 and 185 event days for NICT and SIDC, respectively; M1.0+/0/24 becomes 27 event days for both), they provide the most appropriate approach to enable cross-comparisons. Almost all methods issue multiple forecasts throughout the day; in the course of these comparisons, the forecast issued closest to 00:00 UT was used and others were ignored.

2.2. Standard Metrics and Evaluation Tools

Different performance metrics inform on different performance aspects. This is discussed in Jolliffe & Stephenson (2012) and other references specifically with regards to flare forecasting in Bloomfield et al. (2012), Paper I, Kubo et al. (2017), Steward et al. (2017), and Murray et al. (2018). Hence,

we present a number of metrics and evaluation tools, but for brevity, we refer to any of the above references for the definitions of specific metrics.[19]

Graphical representations of performance are used due to the wealth of information available in a compact form. Reliability plots (also known as attribute diagrams) plot bins of the predicted probability against the observed number of instances in that event frequency bin. A perfect reliability displays points along the $x = y$ line. A perfect forecast is one in which an event is only and always predicted with a probability of 100%; such a service will only have points in the first and last probability bins. Also included in these plots are the climatological rate (event rate) for the testing period (a $y =$ constant line at the event rate for that testing period) and the no-skill line, which is defined as the bisector between the testing-interval climatology and the "perfectly reliable" $x = y$ line. Additionally, we indicate the relative population of the full sample proportion of forecasts within each bin.

Relative (receiver) operating characteristic (curve), or ROC, diagrams are constructed by plotting the probability of detection (POD) versus the probability of false detection (POFD) as a threshold is varied by which a forecast outcome becomes a "yes" forecast. This threshold is commonly referred to as the probability threshold, $P_{\rm th}$, as it is applied to forecast probabilities, but is applied here even though some methods may not strictly produce probabilities. ROC diagrams measure resolution but not reliability. ROC diagrams include the $x = y$ line to indicate no skill; on an ROC plot, perfect forecasts trace the path from (0, 0) to (0, 1) to (1, 1).

Supplementing the graphical evaluation tools are quantitative metrics. Skill score metrics, in particular, compare performance to that of a reference forecast. These are normalized such that perfect forecasts result in a metric of 1.0, and no skill as compared to the reference results in 0.0. The reference forecast may take various forms; the climatology of the testing period or a random forecast is commonly used (Jolliffe & Stephenson 2012), but it may be any other valid forecast method.

The reliability plots can be summarized by the Brier skill score (BSS), the mean-square-error skill score (MSESS) metric for which the reference is specifically the no-skill climatological forecast of the testing period (see Table 1). This metric answers the question: how well did this method do compared to the underlying climatology?

The ROC curves are summarized here by the ROCSS, also known as the Gini coefficient, both of which are related to the AUC but provide more discrimination (Jolliffe & Stephenson 2012; Leka et al. 2018). The ROCSS and Gini coefficient are normalized such that no skill provides a score of 0.0, and perfect forecasts provide a score of 1.0.

Deterministic (or categorical) forecasts can be valuable when preparing forecasts for a particular customer who may require a specified acceptable rate of false alarms, for example, rather than simply a probabilistic forecast. Four additional metrics based on dichotomous (yes/no) forecasts are included: the Appleman skill score (ApSS) uses the testing interval to construct an "across the board" climatology reference forecast (a single reference forecast according to the event day rate in the testing interval), the equitable threat score (ETS) invokes a random forecast, and the Hanssen & Kuiper skill score/Peirce skill score/True skill statistic (here just PSS/TSS) is the difference between the POD and the POFD (see definitions and discussions in Woodcock 1976; Murphy 1996; Barnes & Leka 2008; Bloomfield et al. 2012; Paper I; Kubo et al. 2017; Murray et al. 2017). These metrics are all based on permutations of the "truth table" entries that compare predicted versus observed outcomes and are discussed at length in the references cited above. Additional numeric metrics, such as the proportion correct (PC, also called the rate correct or accuracy) and the frequency bias (FB; Jolliffe & Stephenson 2012), do not compare to reference forecasts per se and may or may not have a similar normalization as required for a formal skill score. The PC metric is common (but can be misleadingly high even for unskilled forecasts in highly unbalanced samples) and the FB indicates systematic over- or underforecasting, a necessary complement to the TSS metric.

A deterministic forecast is produced by imposing a $P_{\rm th}$ for assigning the probabilities or forecast outcomes to yes/no forecasts. This threshold reflects a probability level for an event at which a "real-world" action/no-action decision has to be taken based on, for example, economic losses incurred from one or another type of error. This threshold is then also used for the dichotomous-based metrics (PC, ApSS, ETS, PSS/TSS, and FB) by which that method is evaluated. The performance of a method according to a dichotomous-based metric may vary as a function of $P_{\rm th}$; this is demonstrated in ROC curves where the vertical distance of each point of the curve from the no-skill $x = y$ line reflects the PSS/TSS and thus the method's discrimination between events and non-events as $P_{\rm th}$ is varied (see the discussion in Paper I). Generally speaking, the methods here are either not explicitly optimized for a particular $P_{\rm th}$ during their training or the training method implicitly maximizes a particular metric that effectively optimizes the system at $P_{\rm th} = 0.5$. All but one method produced probabilistic forecasts; for the one that did not (NICT), outputs of 0.0 and 1.0 were assigned "no" and "yes" forecasts, respectively.

Hence, we adopt $P_{\rm th} = 0.5$ to compute dichotomous-based metrics for all methods. A few methods provide custom forecasts to customers with different $P_{\rm th}$ or routinely provide their alerts above a particular $P_{\rm th}$, and those were invited for evaluation with a custom $P_{\rm th}$ (none were submitted). Unless specified otherwise, selecting $P_{\rm th} = 0.5$ for categorical-based metrics is an allowable choice for all methods. All probabilities for all forecast methods accompany this publication (Leka & Park 2019) and are thus available for readers to calculate additional metrics, with $P_{\rm th} \neq 0.5$ for example.

For all methods, missing forecasts were assigned a probability of $p = 0.0$ for that day. This is appropriate for operational forecasts, where missed or skipped forecasts should be penalized. Most operational methods have built in backup sources of data, forecasts, or the ability to forecast prior climatology in the event of, for example, data interruption (see additional details in Paper III).

We do not present the popular "maximum TSS" ($\rm TSS_{max}$) for two reasons. First, an "optimal $P_{\rm th}$" with which $\rm TSS_{max}$ is calculated should be established based on information obtainable only from the training interval, rather than the testing interval itself, as is common practice. No method supplied such a customized $P_{\rm th}$ to use. Determining an optimal $P_{\rm th}$ from which to achieve a maximum TSS score based on testing-period information is not consistent with a purely operational

[19] See also http://www.cawcr.gov.au/projects/verification/#What_makes_a_forecast_good and https://www.nssl.noaa.gov/users/brooks/public_html/feda/note/reliroc.html for broad discussions and numerous definitions.

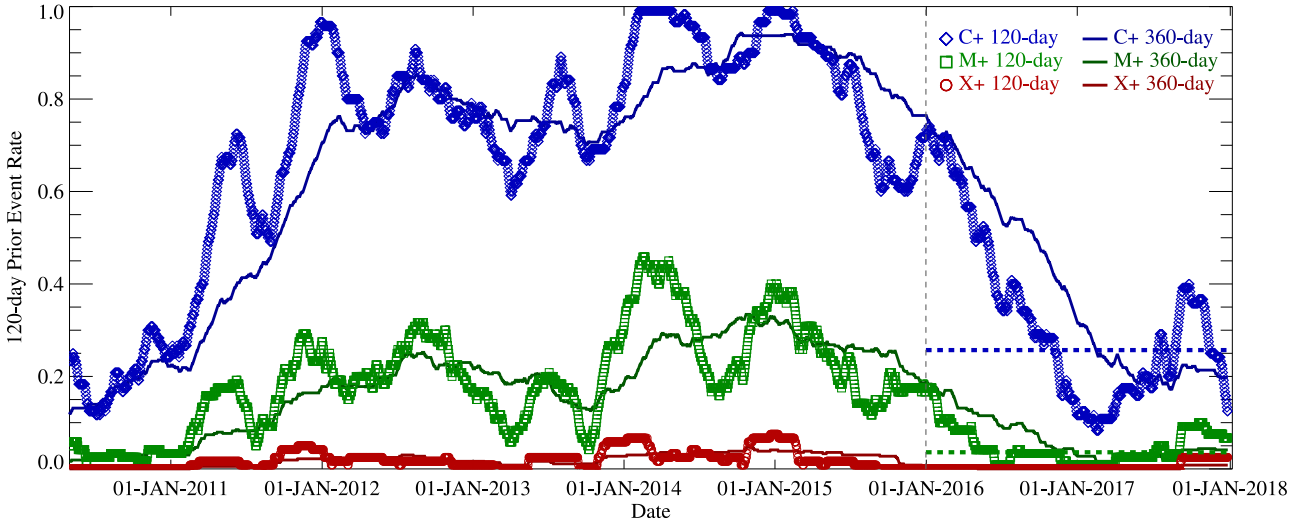

Figure 2. The 120-day prior climatology and 360-day prior climatology are plotted for the `C1.0+/0/24` and `M1.0+/0/24` event definitions, plus the same for an `X1.0+` threshold for completeness, from the start of the *SDO* mission (2010 May 1) through the testing interval, whose start is indicated by a vertical dashed line. The climatological event rate of the testing interval is indicated by horizontal dashed lines over that time period. Each symbol (as indicated) represents the daily full-disk event rate for the prior 120 days (up until but not including the date on which the point falls) as well as for the curves indicating the 360-day prior climatology. The 120-day prior climatology is used as the unskilled reference forecast in the MSESS_clim and ApSS_clim metrics in Figure 5.

approach. The optimal $P_{\rm th}$ can have a correspondence to the underlying event rate (Bloomfield et al. 2012; Paper I), which varies according to the solar cycle and from one cycle to the next as discussed below.[20] Hence, there is limited "actionable information" in determining the optimal $P_{\rm th}$ from a training period for future forecasting. Second, the $P_{\rm th}$ for each method used to achieve $\mathrm{TSS}_{\rm max}$ will differ from each other and will depend on the event definition, so interpreting these results is challenging (see discussion in Paper I). That being said, one can roughly estimate $\mathrm{TSS}_{\rm max}$ for each method from the shape of its ROC plot (i.e., the point of maximum vertical departure from the no-skill $x = y$ line).

2.3. Highlighted Metrics: Comparison against No-skill Operational Forecasts

All metrics discussed thus far explicitly evaluate the performance of forecasts against the outcome of the testing interval. In true operational settings, however, an appropriate reference forecast against which to judge performance is more appropriately the best "unskilled" forecast available (Murray et al. 2017; Sharpe & Murray 2017). In other words, for operational forecasting, it is appropriate to separately and specifically ask: to what extent is the method in question an improvement beyond what would be otherwise available by simply using an unskilled forecast? If a forecasting method cannot perform better than this unskilled forecast, then it does not add any skill or value beyond that unskilled forecast.

To construct a no-skill forecast for day t for the event definition in question, we use an event rate determined over the prior N days up to and including $t - 1$. The resulting event rate is then used as the reference forecast's predicted probability for that date t. We choose $N = 120$ days as suggested by Sharpe & Murray (2017). This unskilled reference forecast does vary, as shown in Figure 2—in particular, decreasing from >0.5 to <0.5 for `C1.0+/0/24` within the testing interval. Its abrupt variation on short timescales (e.g., around 2017 September; see also Figure 5 of Sharpe & Murray 2017) likely reflects active region recurrence patterns and space weather effects rather than reflecting longer-range climatology (see discussion on climatology variations in McCloskey et al. 2018, and the 360-day prior climatology curves also shown here in Figure 2). However, a 120-day prior climatology forecast (CLIM120) avoids significant lag against the fairly rapid event-rate changes that occur at the beginning and end of the solar magnetic cycles evident in the 360-day prior climatology curves. Either provides a valid unskilled forecast and a valid reference forecast for associated metrics, with expected performance differences and resulting scores—as would a no-skill forecast using yet another value for N. The CLIM120 is included for evaluation along with all other methods as a "sanity check" on the performance of this reference forecast.

Two metrics are constructed using this unskilled forecast as the reference. A metric "MSESS_clim" is analogous to the BSS as based on the mean square error (MSE) of the forecast probabilities. However, instead of the testing-period climatology as defined for the BSS, the MSESS_clim uses the prior 120-day event rate ("120-day prior climatology") as the reference forecast. Analogously, we compute an ApSS for which the across the board forecast for any given day is dictated by this reference; the resulting accuracy is computed and used as the reference forecast in the "ApSS_clim" score.

3. The Method Performances

Results are shown here for the metrics and evaluation methodology described in the previous section. Note that if a particular method is highlighted in the text as an example of a particular trend, it will rarely be the only example, and such a callout does not mean other methods are exempt from said trend. Such callouts refer to `M1.0+/0/24` results unless otherwise noted.

First, in Figure 3, the reliability diagrams (attribute diagrams) are shown, comparing predicted probabilities to the observed frequencies across 20 probability bins. The predicted

[20] Some methods (e.g., A-EFFORT) do establish optimal $P_{\rm th}$ levels during training and apply them in order to issue alerts. They elected to not invoke these $P_{\rm th}$ for the evaluations here.

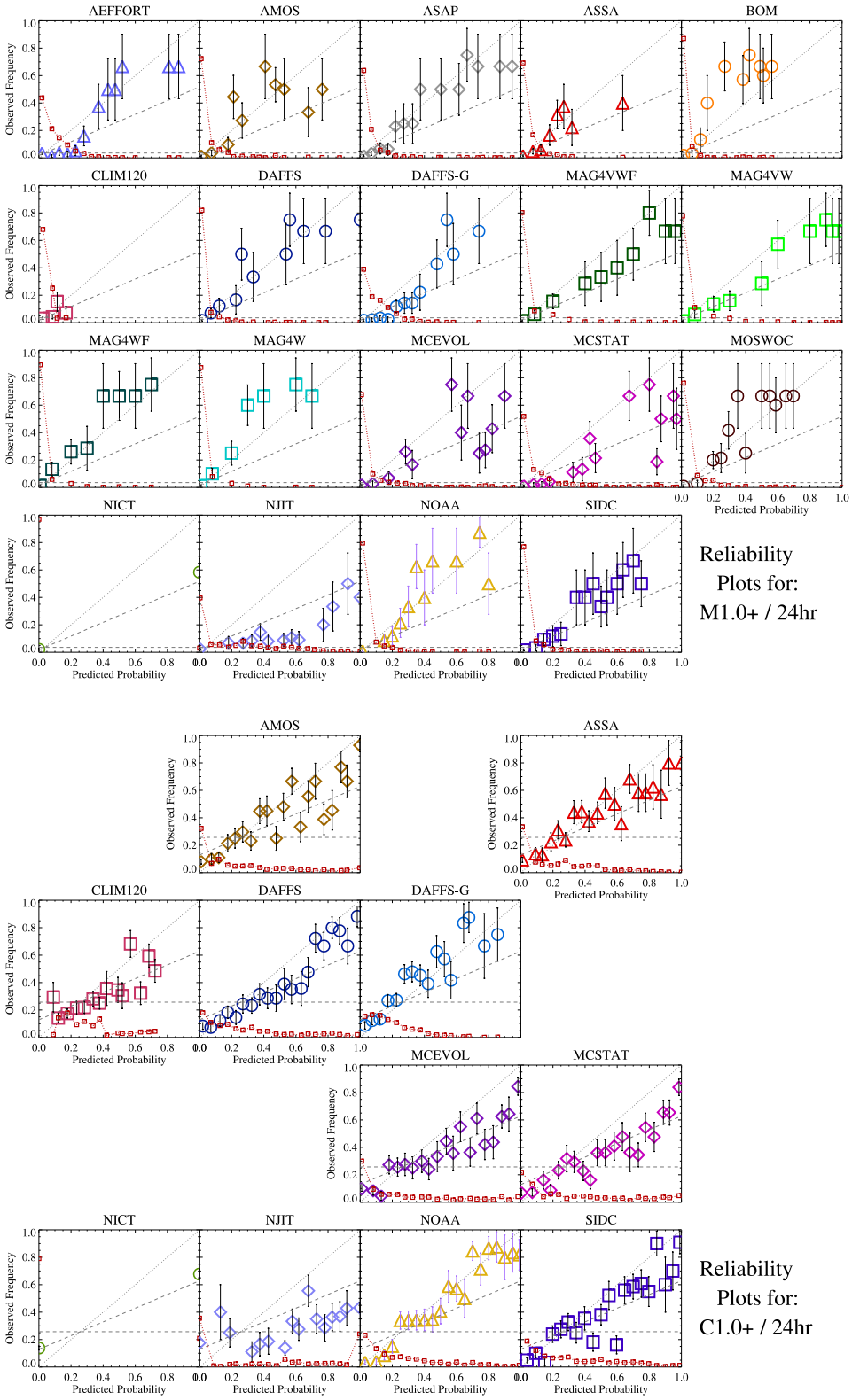

Figure 3. Reliability plots (attribute diagrams) for each method, indicating the performance of the probabilistic forecasts as named: the "$x = y$ perfect reliability" dotted line, the climatology level (horizontal dashed line), and the no-skill line (sloped dashed line) that lies between the two. Additionally shown (the red dotted line and small square) is the fraction of the total sample for which a forecast exists for each bin. Each method has an assigned color/symbol combination (Figure 1), where related methods (e.g., from the same institution) have the same symbols and are plotted with colors in the same family ("nearby" in hue). Results are shown for `M1.0+/0/24` (top) and `C1.0+/0/24` (bottom); fewer methods predict the latter than the former. Results were not calculated for `X1.0+` due to extremely small number of events in the testing interval.

probabilities are indicated on the x-axis by the average of the probabilities in that bin. Points in each bin are accumulated and thus accurately reflect the distribution whether from continuous probabilities or discrete forecast probabilities. This figure also displays the symbol and color schemes devised to both compare methods and intercompare between related methods (e.g., variations from the same institution, see Figure 1). Most methods provided a form of M1.0+/0/24 forecasts (natively, or computed as per Appendix B.2). A subset of methods also produces forecasts for C1.0+/0/24, and those are displayed as well. The decision regarding whether to produce forecasts for these smaller flares rests on the facility or agency according to resources, customer needs, and perceived threat; if publicly available, these forecasts were included. Most methods do provide a forecast for X1.0+; however, the number of events was so small during the testing period as to be uninformative (see Table 1). The error bars are determined by the number of points and events in each bin (Wheatland 2005; Paper I); for a reliability value, R, in a particular bin, $\sigma_R = (R((1-R)/N_{\rm bin}+3))^{1/2}$ with $N_{\rm bin}$ being the number of points in that bin.

The reliability diagrams graphically display trends of overforecasting (see, e.g., MCSTAT) or underforecasting (see, e.g., MAG4W) the issued probabilities. Some methods more systematically perform errors of one type (e.g., BOM), while others display a mix according to the probability bin (e.g., AMOS) but not an obvious dominance of one error or the other. The reliability plots also highlight that some probabilistic methods provide predictions covering the full range of probabilities (e.g., MAG4VWF), while others do not provide predictions at the highest probabilities (e.g., ASSA). The case of NICT, as the sole fully deterministic forecast, appears different due to the assignment of probabilities (see Kubo et al. 2017 for more on evaluation methods for fully deterministic forecasting). This lack of high probability forecasting is more pronounced for larger event magnitude thresholds (e.g., more prevalent here for M1.0+/0/24 forecasts as compared to C1.0+/0/24 forecasts), which is a trend noted in Paper I. Most of the methods here are probabilistic with the exception of the NICT facility, which produces deterministic forecasts. Larger flares are less frequent, and probability-based forecasts will train to reflect that fact, which reduces the presence of high-probability forecast values.

The ROC curves for all methods are presented in Figure 4, using the same color and symbol scheme. The $x = y$ line indicates no ability to discriminate between the two forecast outcomes (forecast for or against an event in the present case). The points on the ROC curve are computed for each distinct probability presented by a method. Hence, methods that provide forecasts in discrete probability bins present fewer points than those that provide continuous-probability forecasts (see, e.g., NICT versus DAFFS). We see a slight increase in the ability of the models that provided forecasts for both event definitions to discriminate for the M1.0+/0/24 results as compared to C1.0+/0/24. This is a generally observed trend (Murray et al. 2017; Leka et al. 2018).

Comparing the reliability versus the ROC plots for a particular method highlights the different information presented by each plot. As an example, the MAG4 results using line-of-sight magnetograms (MAG4W and MAG4WF) versus those using vector magnetograms (MAG4VW and MAG4VWF) appear to show very similar ROC plots while displaying systematically different behavior in the reliability plots (even with different training particulars with regards to longitudinal limitations). Also of interest are the comparative performances of methods that are ostensibly based on the same basic approaches such as Poisson statistics applied to historical region flaring rates (e.g., MCSTAT versus ASSA) or those with human forecasters involved (e.g., NOAA versus MOSWOC).

Figure 5 shows the variety of skill scores and quantitative metrics described in Section 2.2, with approximate 1σ error bars also indicated. There is no straightforward way to estimate uncertainties on the metrics, given the operational approach (e.g., data for a bootstrap evaluation are not generally available). However, we estimate the uncertainties in two ways. First, there are other studies that have employed bootstrap or similar methods to calculate the uncertainties in skill scores (e.g., Bobra & Couvidat 2015; Leka et al. 2018), although the underlying event populations are somewhat different. By adjusting for the smaller sample sizes here, one can estimate a general level of uncertainty in the skill metrics of $\approx$0.06 for C1.0+/0/24, and $\approx$0.10 for M1.0+/0/24. To supplement this estimate, the DAFFS facility (specifically, the magnetic field parameter component) was rerun for the testing interval (2016 January 1–2017 December 31) using a 100-draw (with replacement) bootstrap analysis. Across numerous metrics and variables available in DAFFS, we find the uncertainties range over 0.04–0.09 for C1.0+/0/24 and over 0.05–0.17 for M1.0+/0/24, with the ranges due to whether 1- or 2-variables were tested and the particular metrics used. These estimates are only guidance and do not necessarily reflect the full uncertainty situation. These uncertainties are also likely to be underestimates, because they only account for the random error and no separate bias is calculated for the error estimate itself. For example, when using the full-disk bootstrap, individual days are drawn rather than full-disk passages of individual active regions. Additionally, given the change in the event rate between training and testing intervals, there is likely to be a significant bias present for most methods.

The answer to the question of which methods perform "best" depends on event definition and the metric under consideration. The rank order of performance changes between metrics and between event definitions. This is demonstrated poignantly by MCSTAT/MCEVOL, which score near bottom rank for ApSS but near top rank for TSS/PSS for the M1.0+/0/24 tests.

Some metrics can differentiate performance better than others in these applications. The PC metric for M1.0+/0/24 is uninformative for trying to differentiate between methods due to the large percentage of correct negatives; however, it provides some information for the C1.0+/0/24 analysis. Because the climatology rate does not vary across the 0.5 threshold for M1.0+/0/24, the two Appleman scores (ApSS and ApSS_clim) are identical in this case. In the case of the C1.0+/0/24 event definition, the climatology rate does vary across the 0.5 threshold, and the results for the two scores are slightly different.

That being said, the majority of methods perform similarly to each other—that is, their scores are consistent with each other across metrics. This is particularly the case for the M1.0+/0/24 tests given the estimated uncertainties, although there are arguably performance differences beyond the uncertainties for the C1.0+/0/24 test.

Comparing the reliability plots (based on probabilities) to the FB (which is a dichotomous-based metric employing a single $P_{\rm th} = 0.5$), it appears that the vast majority of methods tend

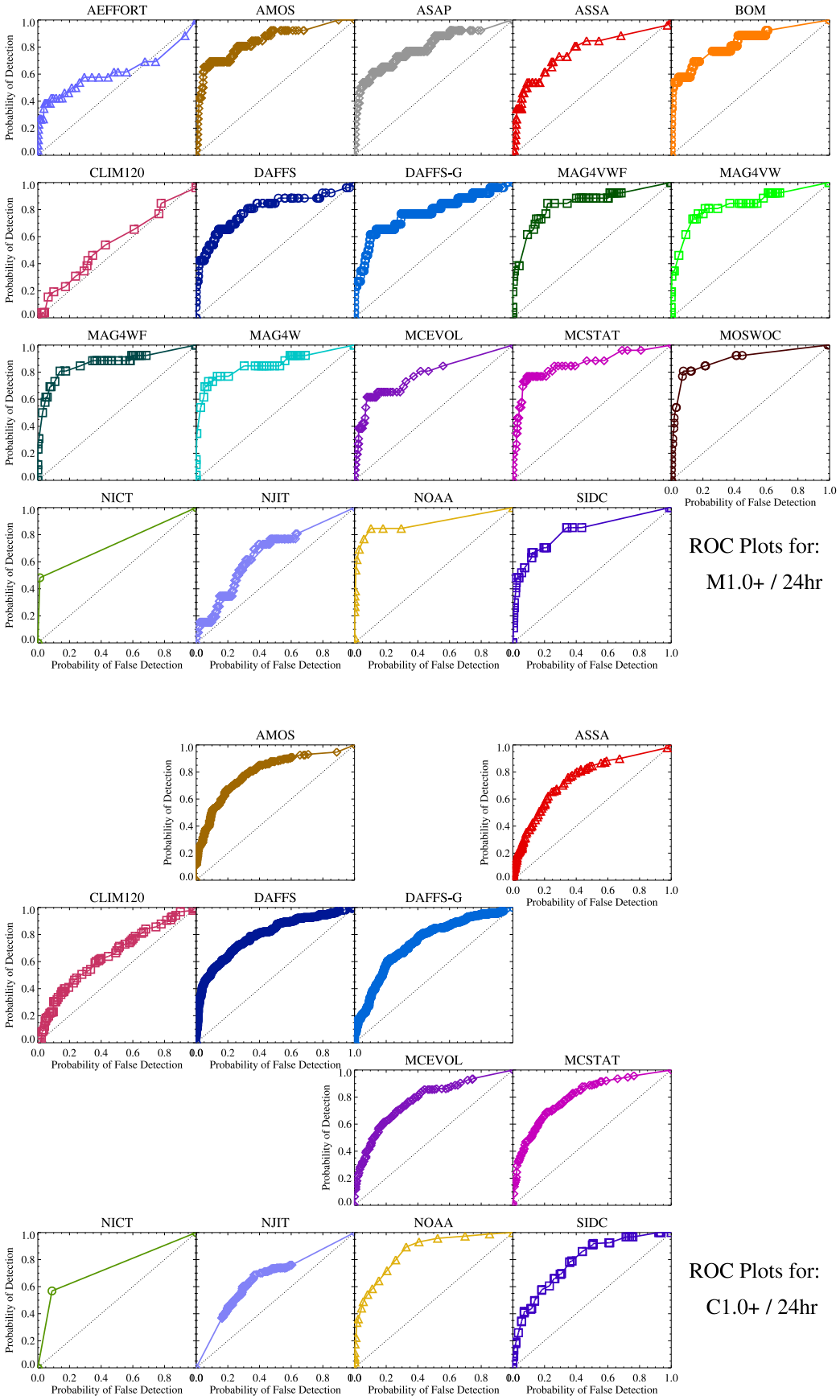

Figure 4. ROC plots with the $x = y$ no-skill line, following the color/symbol scheme of Figure 3.

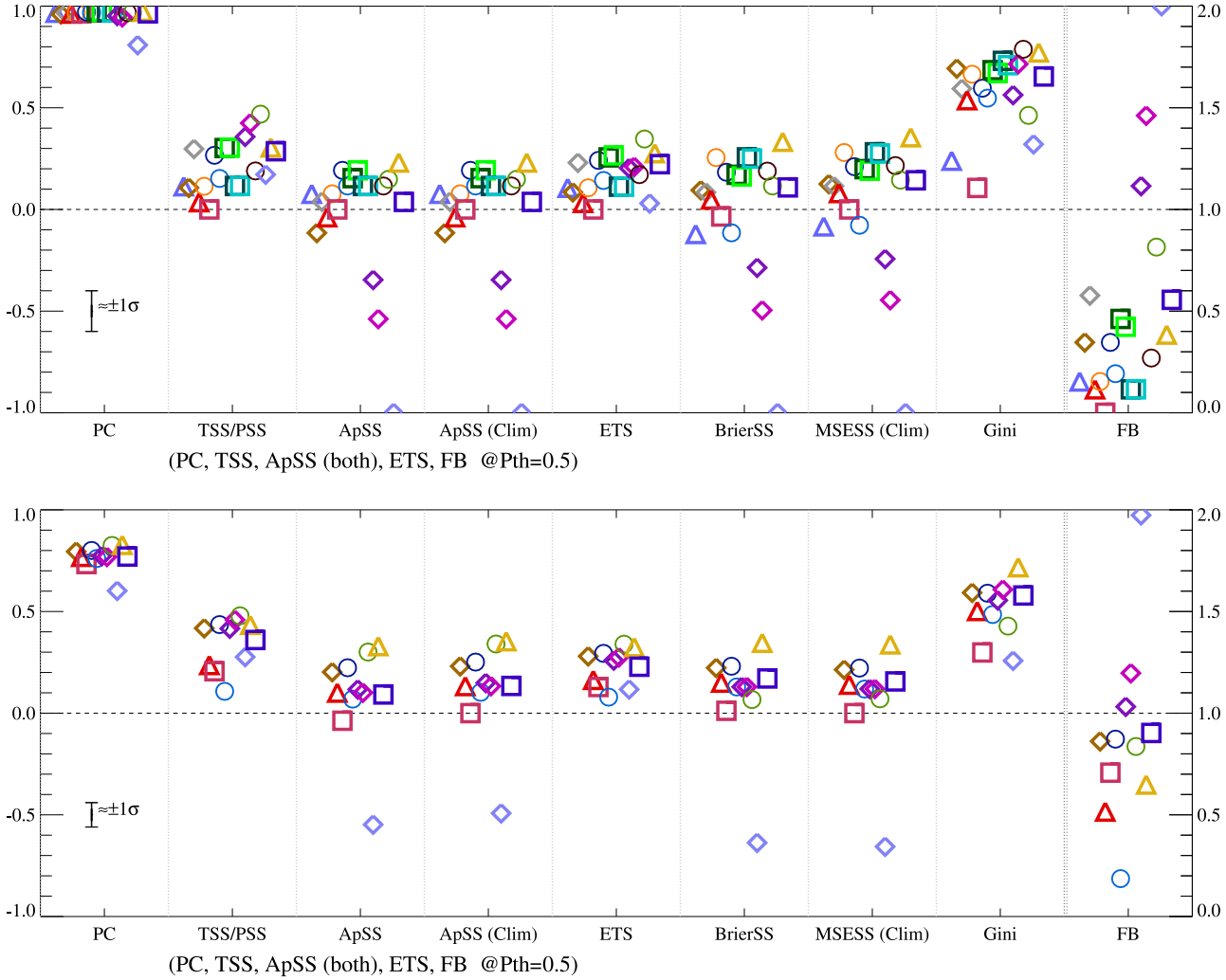

Figure 5. Results from the direct comparison of flare forecasting methods for a variety of performance metrics. (Left to right): the proportion correct, the TSS/PSS, the ApSS (testing period), the ApSS with the 120-day prior climatology reference forecast, the ETS, the BSS, a MSESS with the 120-day prior climatology as a reference forecast, and the Gini coefficient. A lower limit of −1.0 was imposed for the plotting. The final metric is the FB, whose displayed range is indicated on the right-hand axis; a +2.0 limit was placed on this plot. Metrics based on truth tables are calculated using $P_{th} = 0.5$; the BSS, MSESS score (clim), and Gini coefficient are independent of P_{th}. The symbols follow the scheme in Figures 3 and 4 and are offset slightly in the x-dir for clarity in the same order as they appear in Figures 1, 3, and 4). Results are shown for `M1.0+/0/24` (top) and `C1.0+/0/24` (bottom); fewer methods predict the latter than the former. Results were not calculated for `X1.0+` due to the extremely small number of events in the testing interval.

toward underforecasting for larger-flare `M1.0+/0/24` tests by varying degrees (FB < 1.0), with a less pronounced deviation from FB $= 1.0$ for most methods that underwent the `C1.0+/0/24` tests. As mentioned above, the FB score checks the TSS, in that for low event rates that are typical for solar flares, an overforecasting system can attain a high TSS while an underforecasting system is less likely to, so comparing TSS scores should only be performed in the context of an accompanying FB score. As such, for example, confidence in the TSS scores for MCSTAT for the `M1.0+/0/24` test should be tempered somewhat, while the NICT TSS result is more robust.

Different implementations of otherwise the same method can be differentiated and the hoped-for improvements confirmed (or not). The implementations using vector magnetic field data do perform better (albeit only slightly by most metrics) than implementations using B_{los} data within the same general method (e.g., MAG4W* versus MAG4V*, DAFFS versus DAFFS-G). By most metrics, MCEVOL's addition of an evolutionary component to MCSTAT does improve performance, although notably not in the Gini (as visible by the shape of the ROC curve). However, the inclusion of prior flaring history makes almost no difference in performance across the MAG* method (e.g., MAG4W versus MAG4WF, MAG4VW versus MAG4VWF).

None of the operational methods are exceptionally good (i.e., close to 1.0 on any metric, except Gini and PC), although the majority consistently score above no skill for the metrics considered here. Three methods demonstrate arguably poor performance, specifically for the metrics that refer to climatology; these three also show FB > 1.0 (overforecasting). The case of NJIT is fairly well understood and discussed below, while the others will be discussed further in Paper III.

4. Discussion

In this study, we demonstrate two things: first, a methodology to provide meaningful head-to-head comparisons, and second, the present state of operational flare forecasting. With this first direct comparison of forecast methods, benchmarks of

performance by a variety of measures are now provided against which future developments can be tested—an important element of measuring progress in space weather prediction capability.

Regarding the methodology, all forecasting facilities are placed on a level evaluation platform with respect to the full event definition (including thresholds, validity periods, and latencies). Those whose forecasting time differed significantly were afforded custom event lists for evaluation, and those producing both upper- and lower-threshold-limited forecasts were converted to exceedance forecasts to match other methods. Full-disk forecasts ensured that differences in defining solar active regions would not impede the comparisons. The time period chosen was not ideal—it was too short with an arguably very small event list—but in the face of new data sources and a very quiet solar cycle, it was an acceptable and necessary compromise. Most important was how the time period was chosen: a period that was common to all methods that also afforded those methods relying on *SDO*/HMI data an adequate training interval.

The second component of the methodology is the choice of evaluation metrics, and this is arguably a challenge in the context of a direct comparison because it is crucial to ensure that the metrics are all fair (or equally unfair) to all methods. For the presentation here, we select a representative array of dichotomous-based and probability-based metrics, with accompanying graphical evaluation tools, to try and provide as complete a picture as possible. As discussed in Paper I and elsewhere, applying dichotomous-based metrics to probabilistic-based forecasts requires thresholds to be set that may or may not be ideal for a particular method, resulting in unfair penalties. In operational practice, it is challenging to choose the threshold that would ensure optimum performance (by measure of various dichotomous-based metrics) at the time of forecast issuance. As discussed in Bloomfield et al. (2012) and Paper I, an optimum threshold for TSS/PSS is usually close to the climatological event rate, which itself is found only after long-term averages are taken in the testing period. Such information is not available at the time of forecast issuance and may not be optimal for a different metric. For the evaluations here, we encouraged groups to submit deterministic forecasts or to submit probabilistic forecasts and specify thresholds to produce customized deterministic forecasts for particular customers or needs (such as an acceptable error rate of one type or another). None chose to provide other thresholds and thus $P_{\rm th} = 0.5$ was applied to all. As such, we examine how well the methods perform in a deterministic sense if action is only taken when an event is forecast with a probability of 50% or higher.

We make note of metrics that are appropriate specifically for evaluating operational systems, since they specifically query what value the system brings above an available unskilled forecast. The ApSS and BSS, by definition, employ reference forecasts based on the climatology of the testing period but, as discussed, this information is not actionable for improved future performance. We promote evaluations against an unskilled forecast. Here we provide analogous MSESS and an ApSS that employ a 120-day prior climatology as the reference unskilled forecast (as described in Sharpe & Murray 2017), although others may obviously be used. For the testing period herein, the results did not differ substantially from the original version of the metrics. However, the question asked differs in a distinct way, and these metrics are highlighted as part of this work's focus on methodology.

There was not universal agreement in this group regarding evaluation philosophy, specifically with regards to utilizing dichotomous metrics for probabilistic forecasts. The discussion centers on performance variation as a function of the assigned $P_{\rm th}$ in the context of an operational system. While a system may be trained to optimize a particular metric and $P_{\rm th}$, there is no guarantee the performance will be the same with that $P_{\rm th}$ during the testing interval; evaluating a method using a new optimal $P_{\rm th}$ from the testing interval misrepresents the performance when the information needed to assign an optimal $P_{\rm th}$ is unknown at the time of the forecast. One approach for evaluating probabilistic forecasts is to only employ graphical methods, such as the reliability plots and ROC curves, but to also apply metrics, such as the BSS and ROCSS (Gini score), for which no $P_{\rm th}$ is required; this approach is fair (except to the inherently deterministic method(s)) but dismisses some metrics that the community find informative and popular. A second approach is to present all dichotomous metrics in a manner similar to ROC curves, displaying their outcomes as $P_{\rm th}$ is varied and reporting the maximum attained score (with its associated $P_{\rm th}$); but this approach can imply performance better than is attainable in an operational setting and is unlikely to provide guidance for improvement. Hence, the group recognizes that the primary reason for setting a particular $P_{\rm th}$ to apply to probabilistic forecasts is to define a threshold upon which action should be considered according to a particular customer's cost/benefit analysis and resilience against forecasting errors. The full forecast data and evaluation tools used in the present analysis accompany this publication (footnote 20) so that additional metrics—using, for example, a different $P_{\rm th}$—may be calculated by the interested reader.

Regarding the results, generally speaking, no method works extraordinarily well; but we demonstrate that a fair number of methods consistently perform better than various no-skill measures, meaning that they do show definitive skill across more than one metric. No method scores above 0.5 (i.e., halfway between "no skill" and "perfect") across all evaluation metrics, and for a number of metrics no method provides results above 0.5. The specific ordering of performance varies according to the metric and event definition: there is no single "best" method, especially given the estimated uncertainties in the metrics. Among methods that provide different versions, the versions generally behave similarly in some of the gross characteristics (e.g., shapes and sampling for the ROC curves) with subtle offsets reflecting the refinements made between each.

Three particular impacts on forecast method success are worth noting. First, the underlying event rate obviously varies within the solar cycle (Figure 2) and possibly across solar cycles (McCloskey et al. 2018). This will impact the forecasting methods, although the degree of impact will vary depending on training methodology. One example would be that if a method is trained to have high reliability during a time of high solar activity, it may then systematically overforecast during times of declining or lower solar activity. Alternatively, a method may not in fact be particularly reliable during training, but when faced with a particular epoch of the solar cycle (e.g., such as the declining phase with more isolated sunspot groups) it may perform better.

Second, there are always flares, which occur that are not assigned to any particular active region or occur behind the visible limb and may be assigned to a region post facto. During the testing period, there were 41 unassigned C1.0–C9.9 flares and 3 unassigned M1.0–M9.9 flares; in some cases, such unassigned flares were the sole cause of an event day (this is discussed further in Paper IV). Unassigned regions have consequences for training operational systems as well as for evaluating and testing them. The vast majority of methods train on individual regions, and in doing so, they will then underforecast systematically for full-disk forecasts. All region-based forecasting methods will miss days where events are produced by no assigned or detected region.

Third, we can highlight here a distinct case of the impact arising from the lack of a full transition to operational functionality. The NJIT method arguably employs one of the more sophisticated analyses of magnetic field data and shows a distinct skill in the TSS and Gini metrics. However, it arguably performs the worst according to other metrics. Of all the methods, the NJIT system most reflects the research stage of flare forecasting. It was implemented without calibration across a change in instrumentation between training and testing intervals, which, in this case (given the analysis method), could easily cause the systematic overforecasting as evidenced by the metrics. This is an issue faced by many methods in light of aging or changing data sources and the assumed advantage of longer training sets (see Paper III for an additional discussion on that point). Additionally, no provisions were made for issuing forecasts in the event of missing or delayed data, and this severely impacted the metrics in a negative manner. Research methods often report encouraging results, but these must be interpreted in the appropriate context. In parallel, the challenge and effort required to bring research into a fully operational mode to the point that it is ready to undergo evaluation in an operational context must not be underestimated.

From this presentation, it is not possible to further determine why performances differ. Established methods on which national warning centers rely (e.g., NICT versus NOAA) display very different characteristics in the reliability and ROC plots but track fairly well among the evaluation metrics. Newer methods show both improvements and degradation against established ones (e.g., MCEVOL and DAFFS versus MOSWOC and SIDC). However, these differences are fairly subtle (that is, within uncertainties) when examined across all evaluation metrics.

We delve further into the "why" question of performance differences in Paper III by examining the impact of six distinct categories of implementation differences, finding performance advantages to including prior flare information and a human forecaster and performance disadvantages to restricting forecast-relevant data to disk-center observations. We use a novel analysis method to evaluate temporal patterns of forecasting errors of both types (i.e., misses and false alarms) in Paper IV, finding weak support for a hypothesis that including temporal information, such as active region evolution improves a method's ability to successfully forecast, e.g., a region's first flare.

The obvious conclusions from this work are actually broad challenges: new forecasting methods, whether empirical or physics based, need to be evaluated against these established benchmarks with the goals of improved characteristics in reliability and ROC plots and metrics (specifically TSS, ApSS, ETS and BrierSS), all consistently measuring above 0.5 across the full range of event definitions.

We wish to acknowledge funding from the Institute for Space-Earth Environmental Research, Nagoya University for supporting the workshop and its participants. We would also like to acknowledge the big picture perspective brought by Dr. M. Leila Mays during her participation in the workshop. S.-H.P. gratefully acknowledges Dr. Ju Jing for maintaining the NJIT flare forecasting system and providing the archive forecasts. K.D.L. and G.B. acknowledge that the DAFFS and DAFFS-G tools were developed under NOAA SBIR contracts WC-133R-13-CN-0079 (Phase-I) and WC-133R-14-CN-0103 (Phase-II) with additional support from Lockheed-Martin Space Systems contract No. 4103056734 for Solar-B FPP Phase E support. A.E.McC. was supported by an Irish Research Council Government of Ireland Postgraduate Scholarship. D.S.B. and M.K.G. were supported by the European Union Horizon 2020 Research and Innovation Programme under grant agreement No. 640216 (FLARECAST project; http://flarecast.eu). M.K.G. also acknowledges research performed under the A-EFFort project and subsequent service implementation, supported under ESA Contract number 4000111994/14/D/ MPR. S.A.M. is supported by the Irish Research Council Postdoctoral Fellowship Programme and the US Air Force Office of Scientific Research award FA9550-17-1-039. The operational Space Weather services of ROB/SIDC are partially funded through the STCE, a collaborative framework funded by the Belgian Science Policy Office. The authors thank the referees for their constructive comments.

Facilities: SDO(HMI), GONG, *GOES*(XRS).

Appendix A
Operational Forecasting Methods: Additional Details

Here, we list the methods involved in the comparisons. Pertinent details are provided beyond the descriptions provided in the references listed in Figure 1; all times here are quoted in UT. For additional details, we also suggest referring to Paper III, where performance is compared according to specific distinctions.

A.1. A-EFFORT (Academy of Athens, Greece)

A-EFFORT is a space situational awareness (SSA) service of the European Space Agency (ESA), available at http://a-effort.academyofathens.gr/prod/ (with registration). Forecasts are issued at about 00:00 UT and refresh every three hours. Four exceedance thresholds are used: M1.0+, M5.0+, X1.0+, and X5.0+, with a fixed forecast window of 24 hr and 0 hr latency.

There is a single parameter computed from magnetic field data—namely the effective connected magnetic field strength (Beff; Georgoulis & Rust 2007) whose values are translated into probabilities using elements of a Bayesian analysis and Laplace's rule of succession. Beff is calculated directly up to central meridian distances of ±50°; from this limit to ±70°, a magnetic flux-based proxy of Beff is calculated to avoid the impact of severe projection effects.

Each of the four forecasts is computed for each of the active regions present within a solar meridional zone of ±70°, which is identified using a custom active region identification

algorithm (see LaBonte et al. 2007); full-disk probabilities are computed as per Equation (1).

A.2. AMOS (Korean Meteorological Administration and Kyung Hee University)

The Automatic McIntosh-based Occurrence probability of Solar activity (AMOS) model provides daily occurrence probabilities separately for C-, M-, and X-class flares for each NOAA active region and the full disk using McIntosh sunspot group classes and the daily change in the area for the sunspot groups. The details are well described in Lee et al. (2012).

A.3. ASAP (University Bradford, UK)

As described in Colak & Qahwaji (2008, 2009), ASAP also participated in the All Clear workshop in 2009 (Paper I).

A.4. ASSA (Korean Space Weather Center)

The Automatic Solar Synoptic Analyzer (ASSA) system at the Korean Space Weather Center identifies and predicts for a variety of solar activity, including sunspot groups and associated flaring. Flare forecast results are issued hourly at 00:00, with a McIntosh-class-based forecast extending for 24 hr (used here, initiated in late 2013) and a new parameter-based forecast using six major parameters extending for 12 hr. The McIntosh-class-based forecast uses an independent ASSA algorithm (not NOAA determinations) to identify sunspot groups and determines their McIntosh class by estimating their morphological characteristics and produces an independent flaring probability according to the ASSA sunspot–flare archive (not based on otherwise published rates). The ASSA sunspot–flare archive was produced based on statistical matching between ASSA's sunspot group catalog and NOAA's *GOES* soft X-ray events catalog during 1996–2013. A parameter-based method was initiated in late 2016 and provides flare forecasts based on multicomponent linear regression using parameters, such as the number of sunspots in a sunspot group, the total area of sunspots in a group, and the group's longitudinal extent. Unfortunately, forecasts from this second method were not submitted. ASSA forecasts rely on *SDO*/HMI continuum and line-of-sight magnetogram images with no correction for limb-ward effects. Additional details may be found in the user manual (Lee et al. 2013).

A.5. BOM (Flarecast, Bureau of Meteorology, Australia)

The details of the probabilistic model are well described in Steward et al. (2011, 2017). Flarecast II (not yet published but results are submitted here) uses the *SDO* HMI magnetogram imagery analysis capability developed for the original Flarecast model (Steward et al. 2017), plus prior flaring history, and adds a machine-learning technique (logistic regression) to generate a probabilistic forecast. Variables that describe HMI $B_{\rm los}$ magnetograms are selected to minimize Aikake's Information Criteria (AIC), and logistic regression is used to estimate the coefficients of the model and are then used to generate M+, X+, region and full-disk, probabilistic, and categorical deterministic forecasts output for flaring activity over the next 24 hr. In the operational mode, the predictions are updated at 00:00, 06:00, 12:00, and 18:00 UT.

A.6. DAFFS and DAFFS-G (Discriminant Analysis Flare Forecasting System, NorthWest Research Associates (NWRA), USA)

DAFFS is well described in Leka et al. (2018), but it should be noted that it is a fairly "young," recently released system. Note that, being the only method to primarily rely on a quantitative analysis of vector magnetic field data from a non-operational data source (*SDO*/HMI), this method suffered from data problems arising from the data-acquisition mode change that incurred a temporary data misalignment[21] (MAG4V* methods use *SDO*/HMI data in a more limited fashion, see below). The impacted data spanned 2016 April–2017 September and was most damaging for data away from disk center. (The "definitive" data have subsequently been reprocessed; the NRT data will not be.) We noted that it most dramatically impacted some parameters in top-performing combinations but not others. For the results here, we modified DAFFS to run using parameter combinations that performed essentially identically (within the metric error bars) in the training phase but were not as susceptible to the HMI vector data problem: specifically, for the C1.0+/0/24 event definition, the parameter combination was changed to $[E_e, \log(\mathcal{R}_{\rm nwra})]$ and the M1.0+/0/24 event definition parameter pair was changed from what is described in Leka et al. (2018) to $[{\rm FL}_{24}, \log(\mathcal{R}_{\rm nwra})]$.

The DAFFS-G (a tool runs simultaneously and is based primarily on GONG $B_{\rm los}$ data and persistence (NOAA NRT event reports). DAFFS-G is a very "young" release and has not yet been fully optimized for performance. For the forecasts submitted here, the parameter combinations were $[\overline{\nabla(B_z^{\rm pot})}, \Phi_{\rm tot}^{\rm pot}]$ for C1.0+/0/24, and the parameters for M1.0+/0/24 were $[\sigma(\nabla(B_h^{\rm pot})), \Phi_{\rm tot}^{\rm pot}]$, where the "pot" moniker refers to the potential field calculated from the $B_{\rm los}$ data (Leka et al. 2017).

A.7. MAG4* (NASA/Marshall Space Flight Center, USA)

MAG4 is described in Falconer et al. (2011, 2014). This study included four versions:

1. MAG4W: free-energy proxy only using line-of-sight magnetogram.
2. MAG4WF: free-energy proxy and previous flare history using line-of-sight magnetograms.
3. MAG4VW: free-energy proxy only using deprojected HMI vector magnetogram.
4. MAG4VWF free-energy proxy and previous flare history using deprojected HMI vector magnetograms.

MAG4W[F] uses the HMI NRT $B_{\rm los}$ data with no further correction. The MAG4VW and MAG4VWF, like DAFFS, use *SDO*/HMI vector magnetic field data that are, however, only up to 30° from disk center, which were minimally impacted by the data misalignment. In MAG4*F, previous flare information is used, although a region is assumed to be non-flaring if that information is not available.

[21] See http://hmi.stanford.edu/hminuggets/?p=1596 and the *SolarNews* note of 2017 September 1 at https://www.nso.edu/solarnews/20170901/.

A.8. MCSTAT and MCEVOL (MaxMillenium Flare Prediction System, Ireland)

The MCSTAT approach is well described in Gallagher et al. (2002) and Bloomfield et al. (2012), while the MCEVOL approach is well described in McCloskey et al. (2018).

A.9. MOSWOC (Met Office, UK)

The details are well described by Murray et al. (2017). Note that the forecast closest to 00:00 UT was used but is not necessarily the official forecast for that day from MOSWOC, as updates are applied through the (local) night.

A.10. NICT (National Institute of Information and Communications Technology, Japan)

The details of this long-running system are well described in Kubo et al. (2017). The NICT–human approach provides four categorical deterministic forecasts of maximum flare sizes that are unique to the methods: quiet (max: A/B-class), eruptive (max: C−class), active (max: M−class), or major flare (max: X−class). These were converted to probabilities of [0.0, 1.0] for the probabilistic-based analysis and converted to exceedance forecasts.

A.11. NJIT (New Jersey Institute of Technology, USA)

The basic methodology is described in Park et al. (2010). The NJIT method is operational in the sense it produces forecasts automatically but has not been developed further since 2010. It provides probabilistic forecasts of at least one C−, M−, and X−class flare occurrence only for a given NOAA-numbered active region within $\pm 60^{\circ}$ of the disk center; these were converted to exceedance forecasts. The method was trained on 300 primarily flare-productive active regions using *SOHO*/MDI line-of-sight active region magnetic field data in solar cycle 23. However, the forecasts now use HMI line-of-sight data without any crosscalibration between the two data sources.

A.12. NOAA (Space Weather Prediction Center, US National Oceanic and Atmospheric Administration, USA)

The forecasts by NOAA/SWPC have long been considered a standard (Crown 2012) and have set the benchmarks against which methods are measured using the NOAA/SWPC event definitions (see commentary on this in Leka & Barnes 2017). SWPC forecasters begin with a climatological approach. They classify the active regions and assign probabilities according to the historical flaring rates of different sunspot region classes (McIntosh 1990). (Note: SWPC's assignment of active region class is also considered "the standard.") From this, a forecaster may modify a region's probability according to region evolution, flaring trends, and forecaster experience and expertise. These region probability forecasts are combined for a full-disk forecast, which itself may be modified based on flaring history of recently rotated-off regions or indications of a highly active region about to return. Forecasters may also incorporate other model data when available. Initial forecasts issued at 22:00 (the Geophysical Activity Report and Forecast or RSGA) are valid beginning at 00:00 the next day. These are incorporated into the three-day forecast issued at 00:30, with a minimal but not zero probability of a forecast update in the intervening 2.5 hr. Forecasts can, but are not likely to, be updated again before the next three-day forecast is issued at 12:30. The data used in this comparison arise from the three-day forecasts but include the C1.0+/0/24 forecasts that are not generally published.

A.13. SIDC (Solar Influence Data Analysis Centre of the Royal Observatory of Belgium)

The forecaster on duty at the SIDC produces (nominal issue time 12:30UT) a probabilistic forecast each day for the occurrence of X-ray flares over the next 24 hr. Probabilities are provided for flare classes C−, M− and X− separately. A full disk as well as an active region specific forecast are provided. The forecasters use various data sources, the main one being the flaring probability from active regions with the same McIntosh classification. Such probability is then modulated using for example: the specific flare histories for the regions to be forecasted, *SDO*/HMI magnetogram movies, *SDO*/Atmospheric Imaging Assembly (AIA) movies, and *Solar TErrestrial RElations Observatory (STEREO)*/Extreme Ultra-Violet Imager (EUVI) movies, e.g., to assess the flaring activity of active regions rotating onto or off the solar disk. Details on flare forecasting at Royal Observatory Belgium (ROB)/SIDC and its validation procedures are provided in Berghmans et al. (2005) and Devos et al. (2014).

Appendix B
Steps to Produce Full-disk Exceedance Forecasts

B.1. Full-disk Forecasts from Region-based Forecasts

The forecasts considered here are full-disk forecasts, meaning essentially that they treat the Sun as a star. In practice, only one method did not produce full-disk forecasts, meaning that they only provided forecasts for active regions individually. In that case, the region probabilities were combined according to

$$P_{\mathrm{FD}} = 1.0 - \Pi_{\mathrm{AR}}(1.0 - P_{\mathrm{AR}}), \quad (1)$$

where P_{AR} is the probability of an event for each active region, and the product is performed over all active regions for which such a probability is provided. This equation is effectively how all region-forecasting methods produce their baseline full-disk forecasts.

B.2. Class-specific versus Exceedance Forecasts

The results from methods producing class-specific forecasts (e.g., M1.0−M9.9) were converted to exceedance forecasts (e.g., M1.0+ with no upper limit) using conditional probabilities over that method's training interval by the following methodology. Suppose one has the probabilities of occurrence of at least one C−, M−, and X− class flares, respectively, for a given forecast time window, τ, denoted by P(C) for C1.0−C9.9, P(M) for M1.0−M9.9, and $P(\geqslant \mathrm{X1}) = P(\mathrm{X})$ for X1.0+. Then, the lower-bound only probabilities of $P(\geqslant \mathrm{C})$ and $P(\geqslant \mathrm{M})$ can be determined by combining the probabilities of P(C), P(M), and $P(\geqslant \mathrm{X1})$ with their associated conditional probabilities.

The probability of occurrence of at least one flare at the level greater than or equal to M1.0 during τ, (i.e., $P(\geqslant \mathrm{M1})$) can be derived as follows:

$$\begin{aligned} P(\geqslant \mathrm{M1}) &= P(\mathrm{M}) + P(\mathrm{X}) - P(\mathrm{M\ and\ X}) \\ &= P(\mathrm{M}) + P(\mathrm{X}) - P(\mathrm{M}) \times P(\mathrm{X|M}), \end{aligned} \quad (2)$$

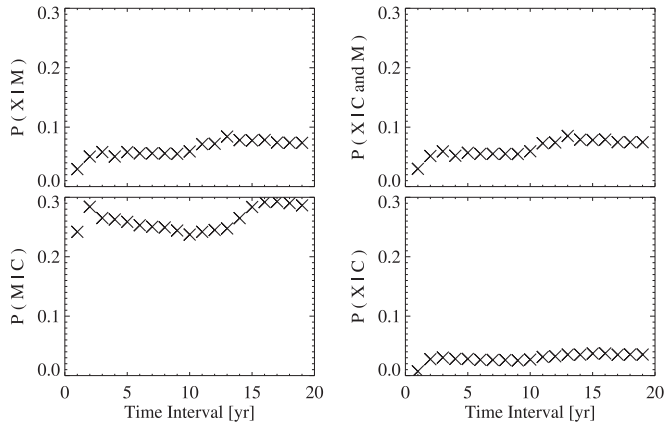

Figure 6. C-, M-, and X-class flare conditional probabilities as function of different time intervals as used for the calculations of exceedance. Time interval extends back in time from 2015 December 31.

where P(M and X) is the probability that both M- and X-class flares will occur at least once during τ, and P(X|M) is the conditional probability of at least one X-class flare occurring given at least one M-class flare occurred during τ.

Similarly, $P(\geqslant$C1) can be determined as follows:

$$\begin{aligned} P(\geqslant \mathrm{C1}) &= P(\mathrm{C}) + P(\mathrm{M}) + P(\mathrm{X}) - P(\mathrm{C\ and\ M}) - P(\mathrm{C\ and\ X}) \\ &\quad - P(\mathrm{M\ \ and\ \ X}) + P(\mathrm{C\ and\ M\ and\ X}) \\ &= P(\mathrm{C}) + P(\mathrm{M}) + P(\mathrm{X}) - P(\mathrm{C}) \times P(\mathrm{M|C}) - P(\mathrm{C}) \\ &\quad \times P(\mathrm{X|C}) - P(\mathrm{M}) \times P(\mathrm{X|M}) + P(\mathrm{C}) \\ &\quad \times P(\mathrm{M|C}) \times P(\mathrm{X|C\ and\ M}), \end{aligned} \tag{3}$$

where P(X|C and M) is the conditional probability of at least one X-class flare occurring given both C- and M-class flares occurred at least once during τ.

The conditional probabilities are calculated using the NOAA/SWPC historical flare event list data and τ as the prescribed validity interval (e.g., 24 hr) starting from 00:00 UT of a given date. In this case, for example, P(X|M) can be determined as follows:

1. During the training interval for a given forecast method, we find the dates D(M) on which at least one M-class flare occurred.
2. From the dates D(M), we determine the subset D(X|M) of dates on which at least one flare at the level greater than or equal to X1.0 occurred.
3. The conditional probability P(X|M) is then the total number of elements in D(X|M) divided by the total number of D(M).

The other conditional probabilities can be calculated in the same way as P(X|M) explained above. Figure 6 shows the conditional probabilities for different time intervals used for their calculations. Note that the end date of all of the time intervals is fixed at 23:59 UT on 2017 December 31. The conditional probabilities do not significantly change as a function of the time interval. Because our goal is to calculate $P(\geqslant$C1) and $P(\geqslant$M1) from the probabilities of P(C), P(M), and $P(\geqslant$X1) that a given forecast method provides, the proper time interval to use for calculating the conditional probabilities is the training interval for that specific forecast method.

Forecasts for flare-class specific probabilities are converted to exceedance forecasts for the following methods: AMOS, ASAP, ASSA, MOSWOC, NICT, and NJIT.

ORCID iDs

K. D. Leka https://orcid.org/0000-0003-0026-931X
Sung-Hong Park https://orcid.org/0000-0001-9149-6547
Kanya Kusano https://orcid.org/0000-0002-6814-6810
Graham Barnes https://orcid.org/0000-0003-3571-8728
Suzy Bingham https://orcid.org/0000-0002-6977-0885
D. Shaun Bloomfield https://orcid.org/0000-0002-4183-9895
Aoife E. McCloskey https://orcid.org/0000-0002-4830-9352
Veronique Delouille https://orcid.org/0000-0001-5307-8045
Peter T. Gallagher https://orcid.org/0000-0001-9745-0400
Manolis K. Georgoulis https://orcid.org/0000-0001-6913-1330
Kangjin Lee https://orcid.org/0000-0001-8969-9169
Vasily Lobzin https://orcid.org/0000-0001-5655-9928
Sophie A. Murray https://orcid.org/0000-0002-9378-5315
Rami Qahwaji https://orcid.org/0000-0002-8637-1130
Robert A. Steenburgh https://orcid.org/0000-0001-8123-4244
Graham Steward https://orcid.org/0000-0002-9176-2697
Michael Terkildsen https://orcid.org/0000-0002-6290-158X

References

Aschwanden, M. J., Crosby, N. B., Dimitropoulou, M., et al. 2016, SSRv, 198, 47
Barnes, G., & Leka, K. D. 2008, ApJL, 688, L107
Barnes, G., Leka, K. D., Schrijver, C. J., et al. 2016, ApJ, 829, 89
Berghmans, D., van der Linden, R. A. M., Vanlommel, P., et al. 2005, AnGeo, 23, 3115
Bloomfield, D. S., Higgins, P. A., McAteer, R. T. J., & Gallagher, P. T. 2012, ApJL, 747, L41
Bobra, M. G., & Couvidat, S. 2015, ApJ, 798, 135
Centeno, R., Schou, J., Hayashi, K., et al. 2014, SoPh, 289, 3531
Colak, T., & Qahwaji, R. 2008, SoPh, 248, 277
Colak, T., & Qahwaji, R. 2009, SpWea, 7, 6001
Crown, M. D. 2012, SpWea, 10, 6006
Devos, A., Verbeeck, C., & Robbrecht, E. 2014, JSWSC, 4, A29
Domingo, V., Fleck, B., & Poland, A. I. 1995, SoPh, 162, 1
Falconer, D., Barghouty, A. F., Khazanov, I., & Moore, R. 2011, SpWea, 9, 4003
Falconer, D. A., Moore, R. L., Barghouty, A. F., & Khazanov, I. 2014, SpWea, 12, 306
Florios, K., Kontogiannis, I., Park, S.-H., et al. 2018, SoPh, 293, 28
Gallagher, P., Moon, Y. J., & Wang, H. 2002, SoPh, 209, 171
Georgoulis, M. K., & Rust, D. M. 2007, ApJL, 661, L109
Hoeksema, J. T., Liu, Y., Hayashi, K., et al. 2014, SoPh, 289, 3483
Hong, S., Kim, J., Han, J., & Kim, Y. 2014, AGUFM, SH21A-4089
Jolliffe, I. T., & Stephenson, D. 2012, Forecast Verification: A Practioner's Guide in Atmospheric Science (2nd ed.; London: Wiley)
Kubo, Y., Den, M., & Ishii, M. 2017, JSWSC, 7, A20
LaBonte, B. J., Georgoulis, M. K., & Rust, D. M. 2007, ApJ, 671, 955
Lee, K., Moon, Y.-J., Lee, J.-Y., Lee, K.-S., & Na, H. 2012, SoPh, 281, 639
Lee, S., Lee, J., & Hong, S. 2013, ASSA GUI User Manual, v. 1.07 (Jeju-do: SpaceWeather), http://www.spaceweather.go.kr/images/assa/ASSA_GUI_MANUAL.pdf
Leka, K. D., & Barnes, G. 2003, ApJ, 595, 1277
Leka, K. D., & Barnes, G. 2017, in Extreme Events in Geospace: Origins, Predictability, Consequences, ed. N. Bu zulukova (1st ed.; Cambridge, MA: Elsevier), 65
Leka, K. D., Barnes, G., & Wagner, E. L. 2017, SoPh, 292, 36
Leka, K. D., Barnes, G., & Wagner, E. L. 2018, JSWSC, 8, A25
Leka, K. D., & Park, S.-H. 2019, A Comparison of Flare Forecasting Methods II: Data and Supporting Code, Harvard Dataverse, doi:10.7910/DVN/HYP74O
Leka, K. D., Park, S. H., Kusano, K., et al. 2019, ApJ, 881, 101
McCloskey, A. E., Gallagher, P. T., & Bloomfield, D. S. 2018, JSWSC, 8, A34
McIntosh, P. S. 1990, SoPh, 125, 251

Murphy, A. H. 1996, WtFor, 11, 3
Murray, S. A., Bingham, S., Sharpe, M., & Jackson, D. R. 2017, SpWea, 15, 577
Murray, S. A., Guerra, J. A., Zucca, P., et al. 2018, SoPh, 293, 60
Nishizuka, N., Sugiura, K., Kubo, Y., et al. 2017, ApJ, 835, 156
Park, S.-H., Chae, J., & Wang, H. 2010, ApJ, 718, 43
Park, S.-H., Leka, K. D., Kusano, K., et al. 2019, ApJ, submitted
Pesnell, W. D., Thompson, B. J., & Chamberlin, P. C. 2012, SoPh, 275, 3
Sawyer, C., Warwick, J. W., & Dennett, J. T. 1986, Solar Flare Prediction (Boulder, CO: Colorado Assoc. Univ. Press)
Scherrer, P. H., Bogart, R. S., Bush, R. I., et al. 1995, SoPh, 162, 129
Scherrer, P. H., Schou, J., Bush, R. I., et al. 2012, SoPh, 275, 207
Schou, J., Scherrer, P. H., Bush, R. I., et al. 2012, SoPh, 275, 229
Schrijver, C. J. 2007, ApJL, 655, L117
Sharpe, M. A., & Murray, S. A. 2017, SpWea, 15, 1383
Steward, G., Lobzin, V., Cairns, I. H., Li, B., & Neudegg, D. 2017, SpWea, 15, 1151
Steward, G. A., Lobzin, V. V., Wilkinson, P. J., Cairns, I. H., & Robinson, P. A. 2011, SpWea, 9, S11004
Strugarek, A., Charbonneau, P., Joseph, R., & Pirot, D. 2014, SoPh, 289, 2993
Wheatland, M. S. 2000, ApJL, 536, L109
Wheatland, M. S. 2005, SpWea, 3, 7003
Woodcock, F. 1976, MWRv, 104, 1209