During the refereeing process, there was discussion regarding potential problems for the validation statistics introduced by using 8hr sampling and 24hr validity periods.

Below we include some of the points made by both the reviewer and the authors, edited slightly for clarification & confidentiality, for any interested reader's edification.

Reviewer (edited):
The concern identified is in the combination of forecasts being used for the forecast verification metrics. Overlapping forecast windows are used, (see Table 1), indicating that there are three 24-hr forecasts per day. This is not how forecast verification is meant to be performed, because the statistics behind the combination of forecast-observation pairs in standard verification metrics assumes independent and identically distributed (i.i.d) - i.e. each forecast-observation pair is independent of every other pair and they follow the same probability distribution. Overlapping forecast windows lead to non-independent forecast-observation pairs because the observation of a single flare affects the interpretation of multiple overlapping forecasts. By choosing just one of your current forecast times (e.g. 00:00 UT) and generating the forecast verification metrics, reliability diagrams, and ROC curves for that particular set of edge-to-edge forecast windows (i.e. no overlapping occurring at all), they will then have the benefit of being statistically correct in their application and in your subsequent interpretation of their values and behaviors.

Authors' Response (edited):
[…] The AFT maps are evolved continuously (at a 15min cadence, with hourly data assimilation). We *sample* this evolving system every 8 hours at 00:00, 08:00 and 16:00 and at each time the parameters are computed independently from the other times. As such, we argue that in fact each sampled time comprises an independent forecast–observation (or, more precisely a classification-observation) pair. We agree and acknowledge that one could argue there is minimal evolution in 8hr but the AFT developers have incorporated significantly higher cadence evolution based on extensive testing and experience, thus there is reason to believe that there is, in fact, substantive evolution. The value of a computed parameter (e.g. total magnetic flux) at time t and its label (*event / no-event*) vs. the value of the same parameter at time $t+8$ and its label are both valid. The value of the two parameters are not the same but sample the probabilities constructed by all such parameter-label pairs. The labels for the parameter samples at t and $t+8$ may refer to the same event but are binary with regards to the event definition, rather than continuous. Since the values of the parameters are independent, the classification–observation pairs are also independent.

Reviewer (edited):
I understand and agree that the observations are independent of each other, but it is actually the event-observation *label* that I am referring to (i.e., the occurrence -or not- of a GOES flare above the event level definition occuring in the forecast window) and the corresponding independence of these event labels when combined with their associated classifications to generate the truth tables.

As an example, consider the case when one GOES M2.0 flare occurs at 23:00 UT on a particular day. This single flare will cause three overlapping 24hr forecast windows to be labeled as a yes-event. If all the forecast-classifications for those three windows are the same, then combining the results of the overlapping windows into one truth table leads to just one quadrant of the table having 3 added to it. If this always happens then there is no difference in the verification metrics calculated from a single window's truth table and those calculated from the combined three windows' truth table because this factor of 3 cancels from all four quadrants.

However, if there are differences between the forecast-classifications for the three windows that overlap the single M2.0 flare the combined truth table will populate both a correct-forecast quadrant (e.g., TP increases by 2) and an incorrect-forecast quadrant (e.g. FN increases by 1). Then, over the whole testing set the four values in the 2x2 truth table are not independent because the occurrence of each individual GOES flare has the potential to influence more than one quadrant, with no knowledge of how many times this does/doesn't happen or what biasing consequences result. This in turn compromises the validity of many verification metrics, especially those that are based on reference forecasts of random sampling (e.g. Heidke Skill Score and Gilbert Skill Score) or climatology, and will introduce bias between the correct/incorrect forecast quadrants of a given event-label that can compromise even metrics that are not sensitive to climatology (e.g. Peirce's Skill Score).

Although the example provided above relates to truth tables, similar issues will arise for probabilistic forecast verification metrics and reliability diagrams. For example, the Brier Skill Score accounts for the event climatology which will be numerically different for the case of overlapping forecast windows compared to separate end-to-end forecast windows.

Authors' Response (edited):

To address this concern, please see below (1) further analysis, and then (2) additional thoughts on independent labels, and finally (3) our actions with regards to this paper.

(1) Further analysis.
(a) Event Rates and Skill Scores
We performed the NPDA analysis on the data separated by map time, so that there are 3 separate sets of skill scores. Of note, since the sample sizes generally decrease by a factor of 3, the uncertainties in the skill scores (which we estimated) increase by at least that factor (although please note that all of the skill scores result after cross-validation). The results, compared to the "combined" (presented in the paper), are shown below:

Experiment	Map time	Event Rate	Brier Skill Score	Peirce Skill Score (not in paper)
C1.0+, 24hr, "F11"	00:00	0.254	0.262	0.467
	08:00	0.253	0.239	0.452
	16:00	0.250	0.250	0.463
	AVERAGE	0.252	0.250	0.461
	COMBINED (in paper)	0.252	0.264	0.472
M1.0+, 24hr, "F11"	00:00	0.043	0.131	0.576
	08:00	0.043	0.146	0.586
	16:00	0.043	0.136	0.587
	AVERAGE	0.043	0.138	0.583
	COMBINED (in paper)	0.043	0.178	0.594

(We also include the Peirce Skill score, which for reasons surrounding event rates, etc., we discuss but hesitate to include in the paper. As an example dichotomous score we include it here, with Pth = Event Rate as used for the "Δ[FP, FN]" analysis in the paper.)

The results are the following:

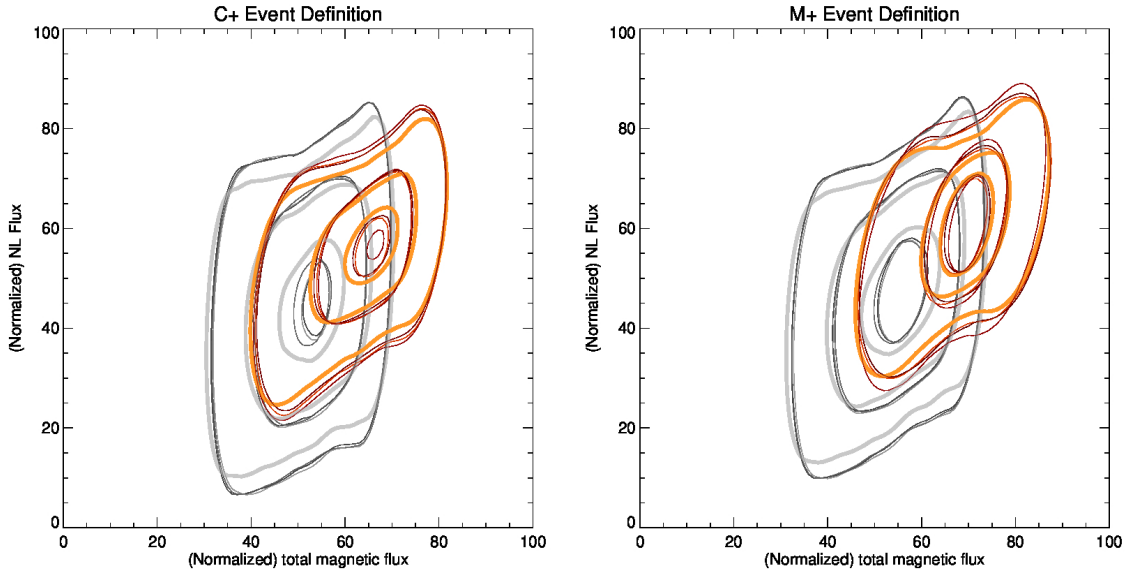
- The event rates for the three times individually average out to the same as the combined (as originally presented), for both C+ and M+ thresholds. This is as expected, since the splitting-data analysis represents sub-sampling and the Sun should be Earth-based agnostic for this experiment.
- The skill scores (BSS and PSS, both) vary between the three times, but are within the expected uncertainty levels (≈0.02 - 0.03, for the reduced sample sizes).
- The average of the three time-separate skill scores are lower, slightly, than those for the "combined".

This is a small experiment. The sample sizes are small and the tests performed above are not exhaustive. Still, there are no substantial differences in the event rates or the skill scores between the time-separate tests, and the only consistent result (that the "combined" results are slightly better than any individual time's results) is easily attributed to a larger sample size and thus, better-defined probability density estimates and hence (slightly) better classification.

(b) Probability Density estimates.

We compared the nonparametric density estimates of the 3 times, separately, and also compared these to the PDEs of the combined dataset, to look for bias or systematic differences (below).

We conclude that for this experiment, there is no difference between the three times, and no significant difference between the single-time PDEs and the combined-



Probability density estimates for no-event (grey) and yes-event (orange), for the three times, matched in hue [00:00 is lightest], event-thresholds as labeled, only on-disk data. The PDEs for the combined dataset are also shown in 8x thicker and lighter-hued lines. Yes-event contour levels:[0.01, 0.065, 0.13], no-event contour levels:[0.01, 0.03, 0.06]. Beyond what might be expected by the overall event rate variations shown in the table above, there are no distinguishing differences or systematic biases in the density functions between the three times or the three times vs. the combined dataset.

time PDEs. The data underlying the two event definitions is the same, of course, so the strong similarities between the event definitions is also as expected; it is primarily the probability levels that change.

(2) Additional thoughts on labels.
I (KD Leka) do think these are interesting points with regards to ensuring that the labels are independent. I acknowledge that there are often counter-intuitive impacts or unintended consequences of sub-sampling, over-sampling, introduced (or removing) imbalance, etc. A similar concern as the referee raises could be applied to the labels in the situation alluded to, for example when the event definition is for “M1.0+ over a period of 24hr”, since it could be argued that there is not an independent label, or classification-observation pairing when more than one event occurs during the validity period. By changing the start time of the validity period (from, say, 00:00 to 12:00), a similar argument could be made regarding biases, changes in event rates, and the resulting table entries. However, the Sun does not know what time it is on Earth, and such re-sampling should not impose bias (but that may not be guaranteed).

(3) What we did for this paper:
This paper is attempting to be “lean, to the point, and a proof-of-concept” while also being statistically appropriate. As such, we believe that it would significantly muddle the analysis to either present in the paper the results of separate analysis for all 3 times, or to only choose one sample time but then face less-than-1σ results for the statistical analysis of the metrics-based impact due to limb events (given the dominance of the on-disk events).

Thus, we have added a paragraph’s worth of text in the Discussion regarding the possible bias due to a lack of independence for the event labels, describing briefly the experiments done to test the introduction of bias from using the over-sampled approach, and state that we find no significant bias. We additionally point out at the end that the data by which interested readers can confirm our tests is available through the associated data repositories.

We will include the table and plots shown above, along with explanatory text, in the Harvard Dataverse repository.