Solar Flare Statistics Don’t Follow Poisson: Overdispersion from Overlapping Active Regions

Joshua D. Ingram[1,2]
Vinay L. Kashyap[2], Bernhard Klingenberg[3], Xiao-Li Meng[4]

[1]Carnegie Mellon University, Department of Statistics and Data Science
[2]Center for Astrophysics | Harvard & Smithsonian
[3]New College of Florida
[4]Harvard University, Department of Statistics

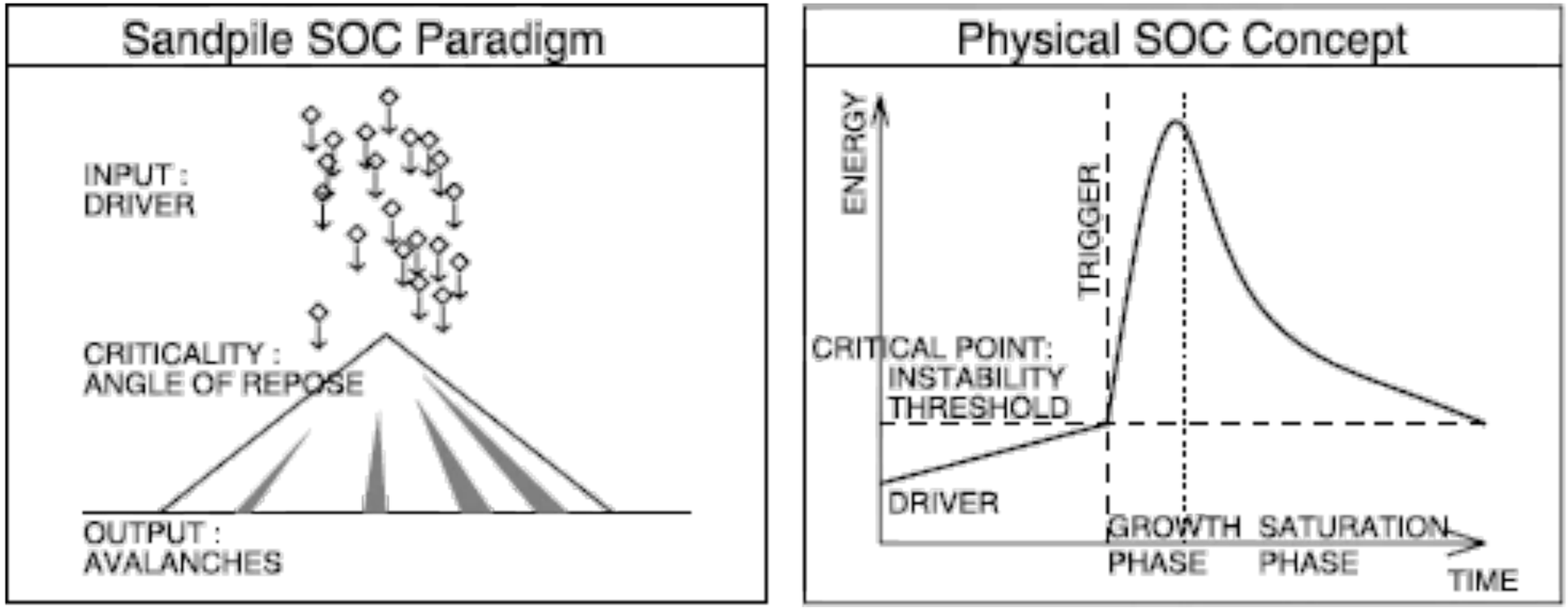
What to Expect

- **Background:** Self-organized criticality (SOC), the Poisson process, and flares
- **Result 1:** Aggregate flare count distributions are overdispersed
- **Result 2:** Flare waiting times within individual active regions depart from the exponential distribution and are overdispersed
- **Result 3:** The observed counting process—a mixture of temporally overlapping active regions—results in overdispersion
- **Moving Forward:** Implications for flare studies and future work

Background:
Self-Organized Criticality, the Poisson Process, and Flares

Self-Organized Criticality and Flares

- Flares modeled based on **self-organized criticality**:
  - **Avalanches**: energy build up and small perturbations in system can trigger cascading events, reaching instability threshold
  - Solar corona evolves to critical state where flares (energy releases) occur across all scales



[1](Figure 1) Aschwanden, M. J. et al., (2014). 25 Years of Self-Organized Criticality: Solar and Astrophysics.

Key Flare Properties for SOC

- Under certain SOC models, we expect:
  - **Energies (*E*) follow a power-law**[1]:

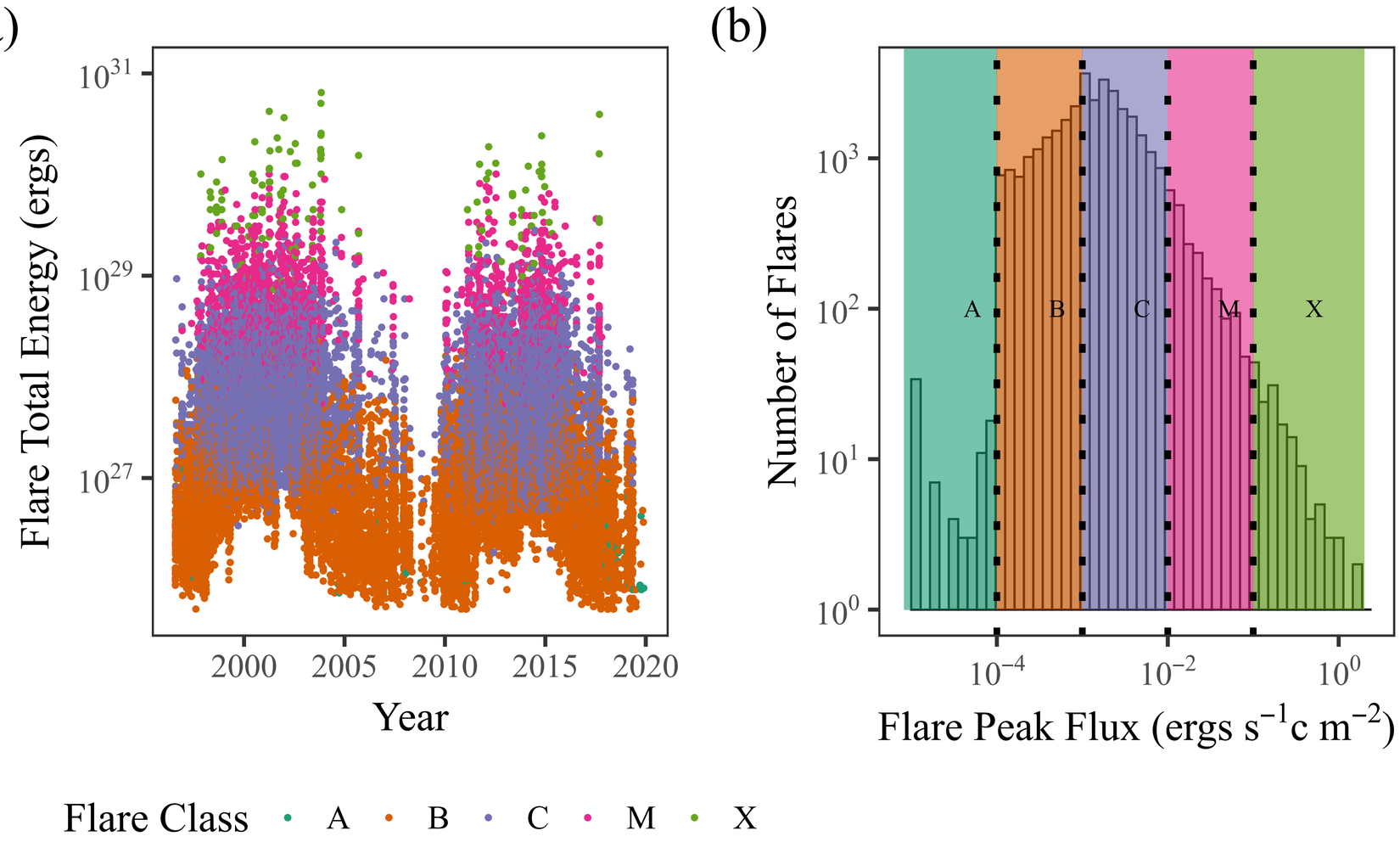
$$N(E) \propto E^{-\alpha}$$

  - **Flare occurrences follow a Poisson process**
    - Counts distributed as Poisson:

$$N(t) \sim Poisson(\lambda t)$$

    - Waiting times ($W_i = T_i - T_{i-1}$) distributed as exponential

$$W_i \sim Exponential(\lambda)$$



[1] Lu, E. T. & Hamilton, R. J. (1991). Avalanches and the Distribution of Solar Flares.

Stationary Poisson Process

- A **stationary Poisson process** is a special case of a counting process, with the properties:

  1. Disjoint intervals are independent

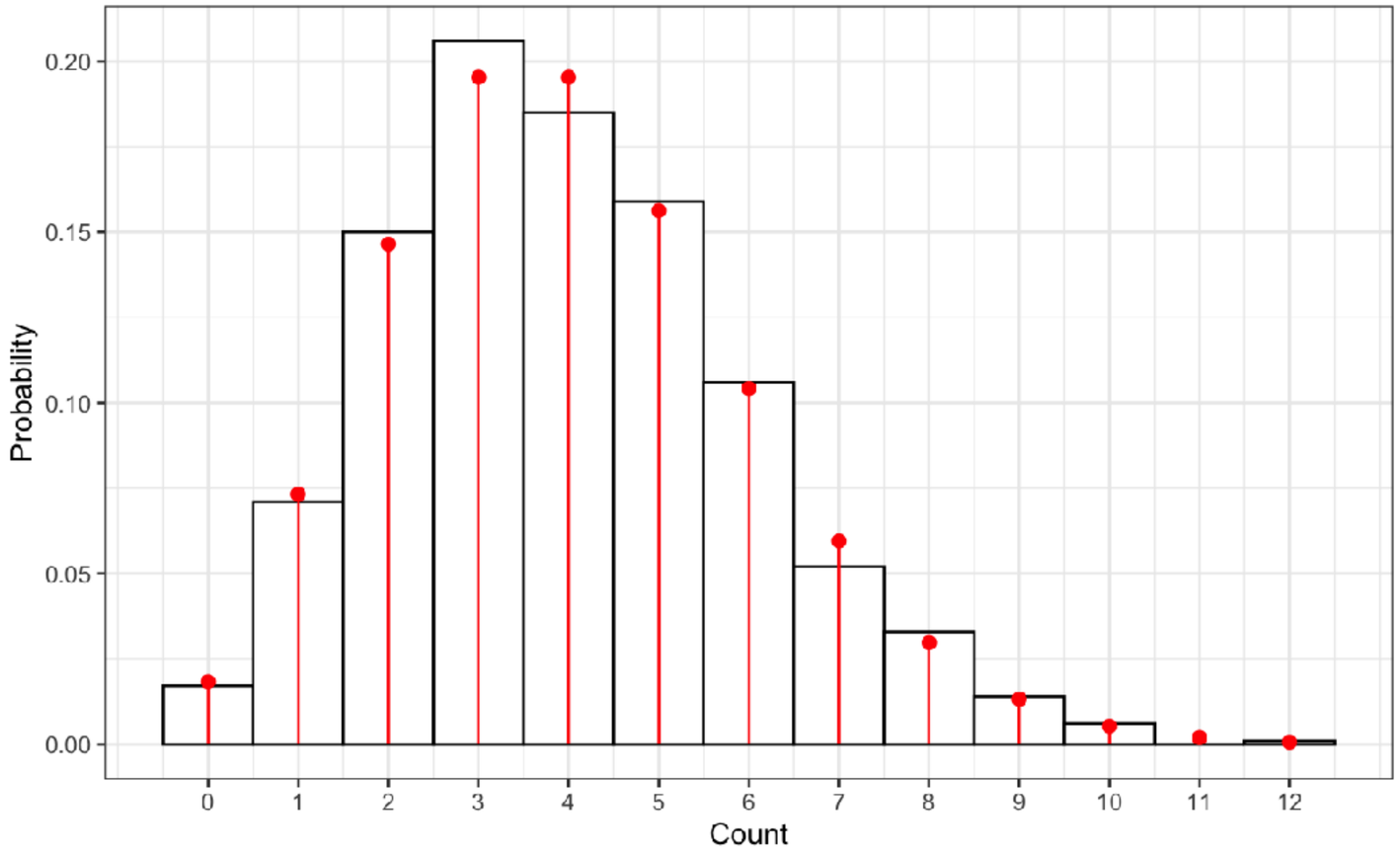
  2. Stationary (constant) rate $\lambda > 0$ for fixed interval

- The distribution of counts follow a **Poisson distribution**. For any $t \geq 0$:

$$N(t) \sim \text{Pois}(\lambda t), \quad P\big(N(t) = n\big) = \frac{(\lambda t)^n e^{-\lambda t}}{n!}$$

- We also have that: $\mathbb{E}[N(t)] = \text{Var}[N(t)] = \lambda t$

- **Solar and stellar flare counts are often assumed to be Poisson distributed!**[1]

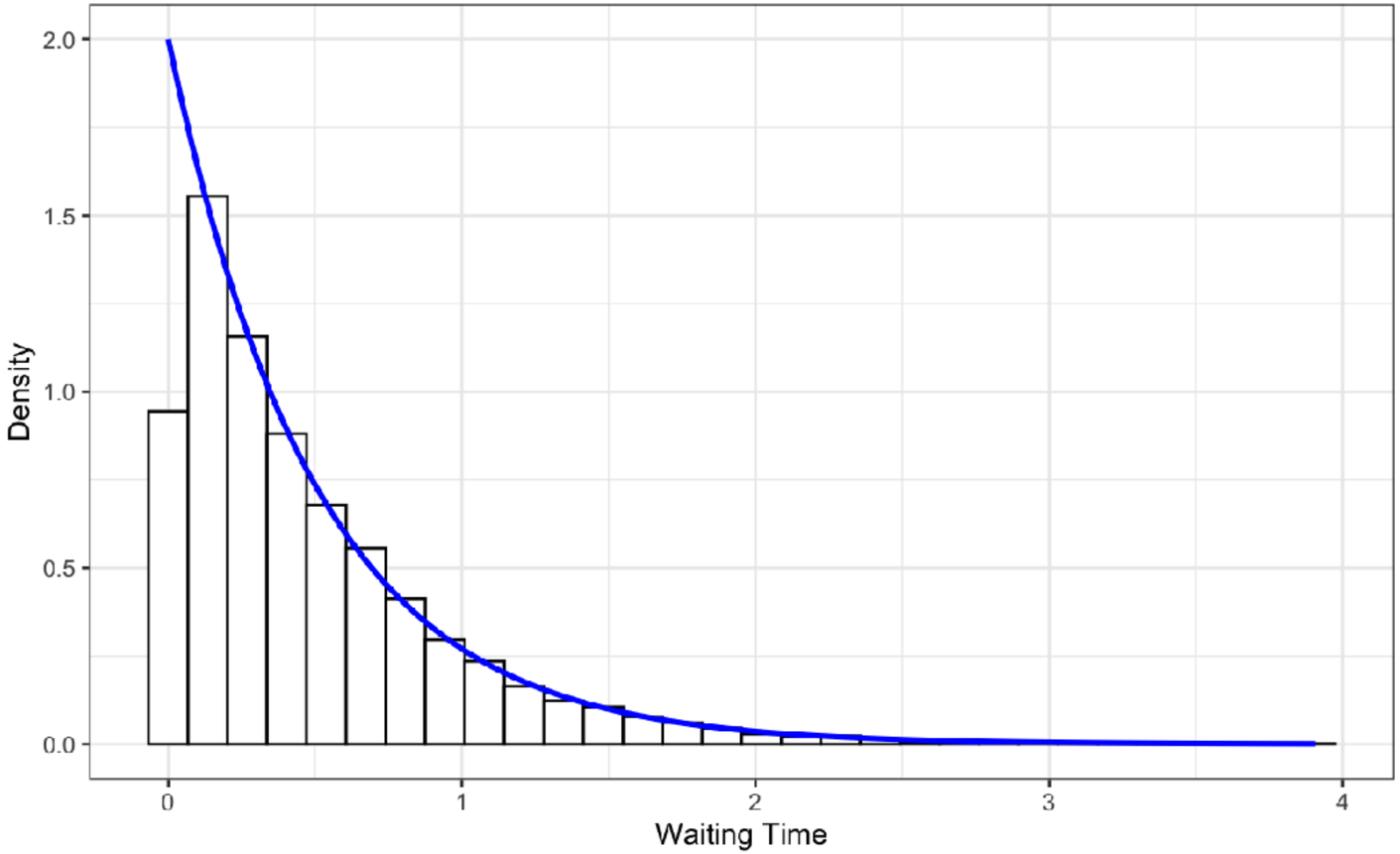


[1] Burton, K. et al. (2025). The Proxima Centauri Campaign - First Constraints on Millimeter Flare Rates from Alma.

Waiting Time Distributions

- The **waiting time** (or inter-arrival time) between two events is defined as: $W_i = T_i - T_{i-1}$
- The waiting times follow an **exponential distribution**:

$$W_i \overset{\text{iid}}{\sim} \text{Exp}(\lambda), \quad f(w) = \lambda e^{-\lambda w}, w > 0$$



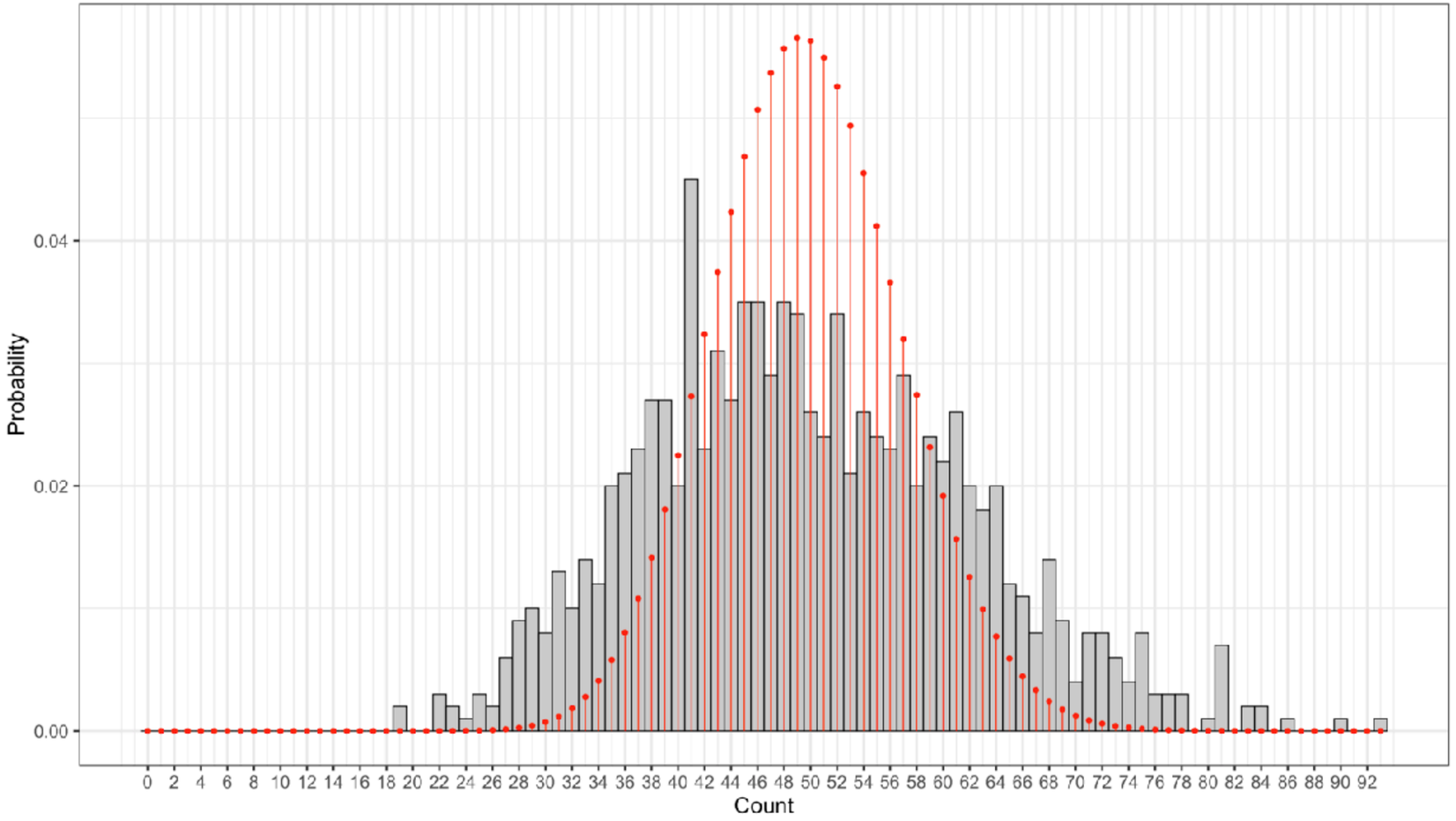
Non-Stationary Poisson Process and Overdispersion

- **Overdispersion** describes when:

  $$\mathrm{Var}[N(t)] > \mathbb{E}[N(t)]$$

  relative to the Poisson distribution

- Overdispersion can be caused by a **non-stationary Poisson Process**. Examples include:
  - A deterministic, time-varying rate $\lambda(t)$
  - Cox process, where $\lambda(t)$ is stochastic



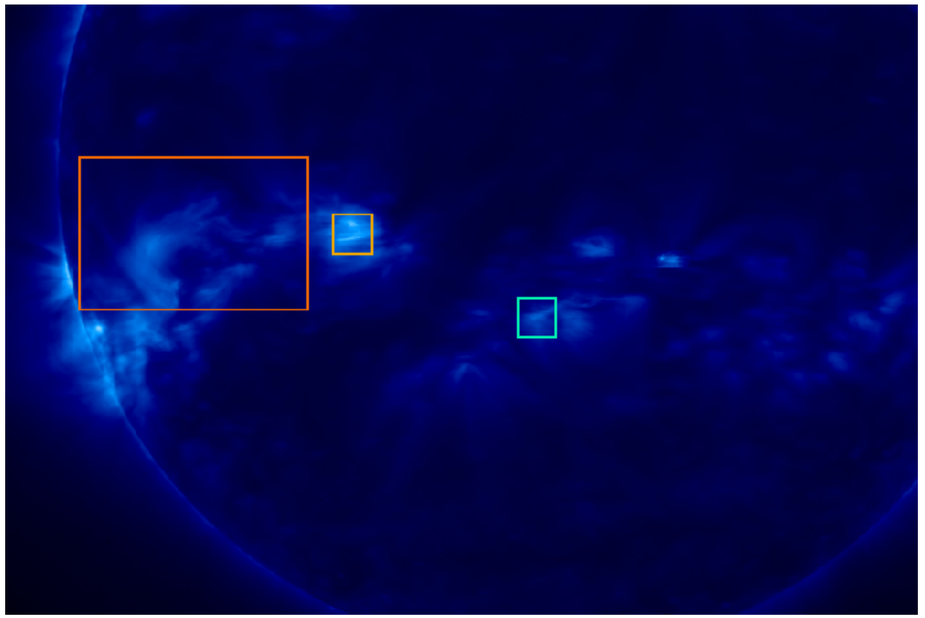
Prior Results on Flare Waiting Times

- Several studies have identified **power-law behavior of flare waiting times**, rather than exponential[1,2]
- Power-law behavior explained as a consequence of a **non-stationary Poisson process**. Approaches:
  - Bayesian blocking methods to fit **(discrete) piecewise constant** exponential distributions[3]
  - Polynomial functions for **continuously-varying rate parameter**[4]

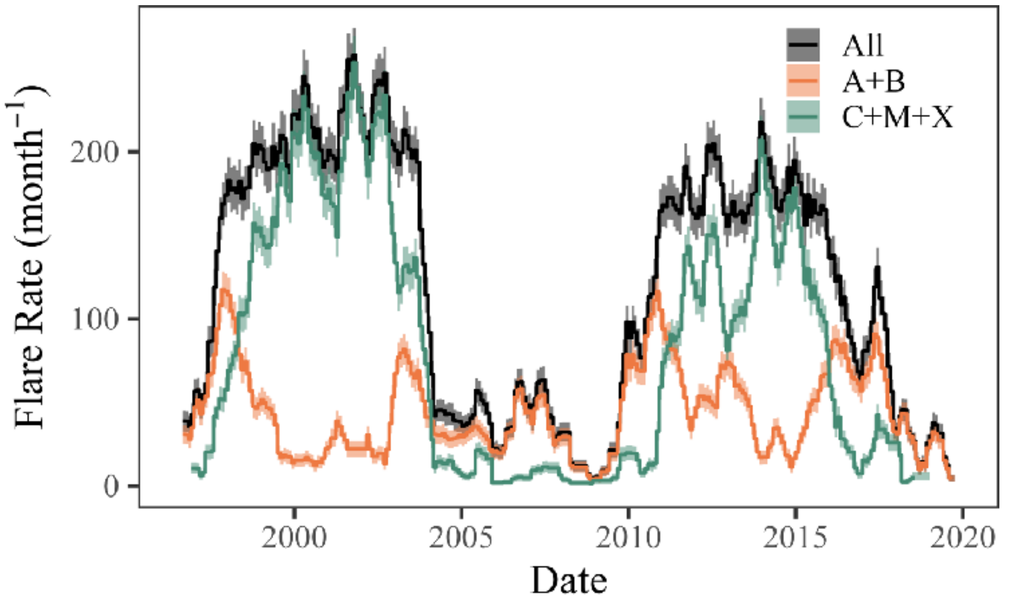
[1]Wheatland, M. S. et al. (1998).; [2]Bofetta, G. et al. (1999).; [3]Wheatland, M. S. (2000).; [4]Aschwanden, M. J. et al. (2021).

Why, and What About Counts?

- Why the power-law/non-stationary behavior?
  - Solar cycle?
  - Sympathetic flaring?
  - **Unresolved, latent subprocesses?**
- What about count distributions?
  - **Should we still assume a Poisson distribution for solar and stellar flares?**

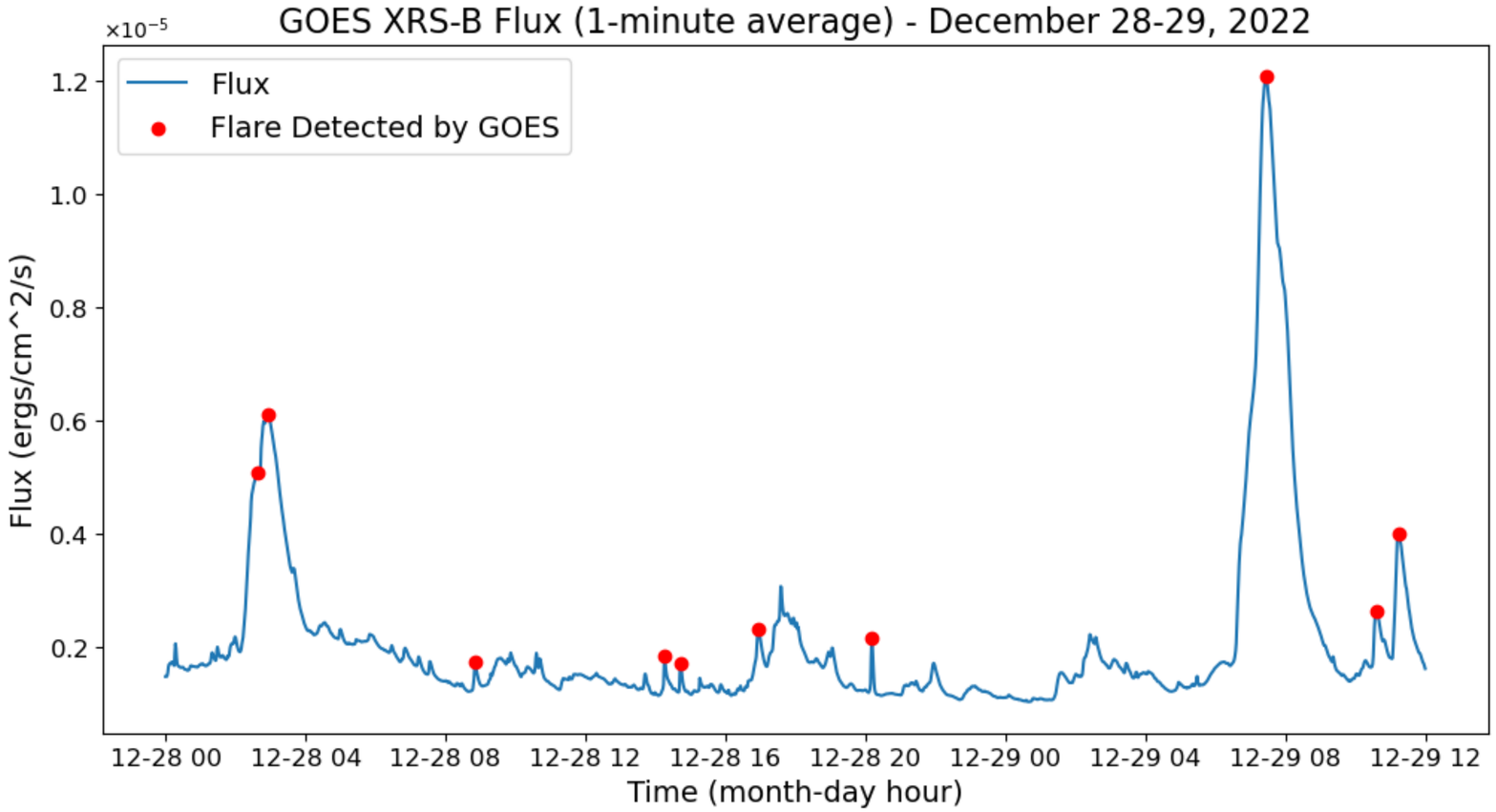


Credit: NASA/SDO



Data for our Analysis

- GOES flare database generated from detections using X-Ray Sensor (XRS) 1-minute average flux in 1-8 Å (XRS-B) passband
- Includes flares from July 1996 — December 2019 (Cycles 23 and 24)
- Properties in catalog:
  - start, peak, and end times
  - Peak flux and total energy
  - (Some) Solar disk locations and active region assignments



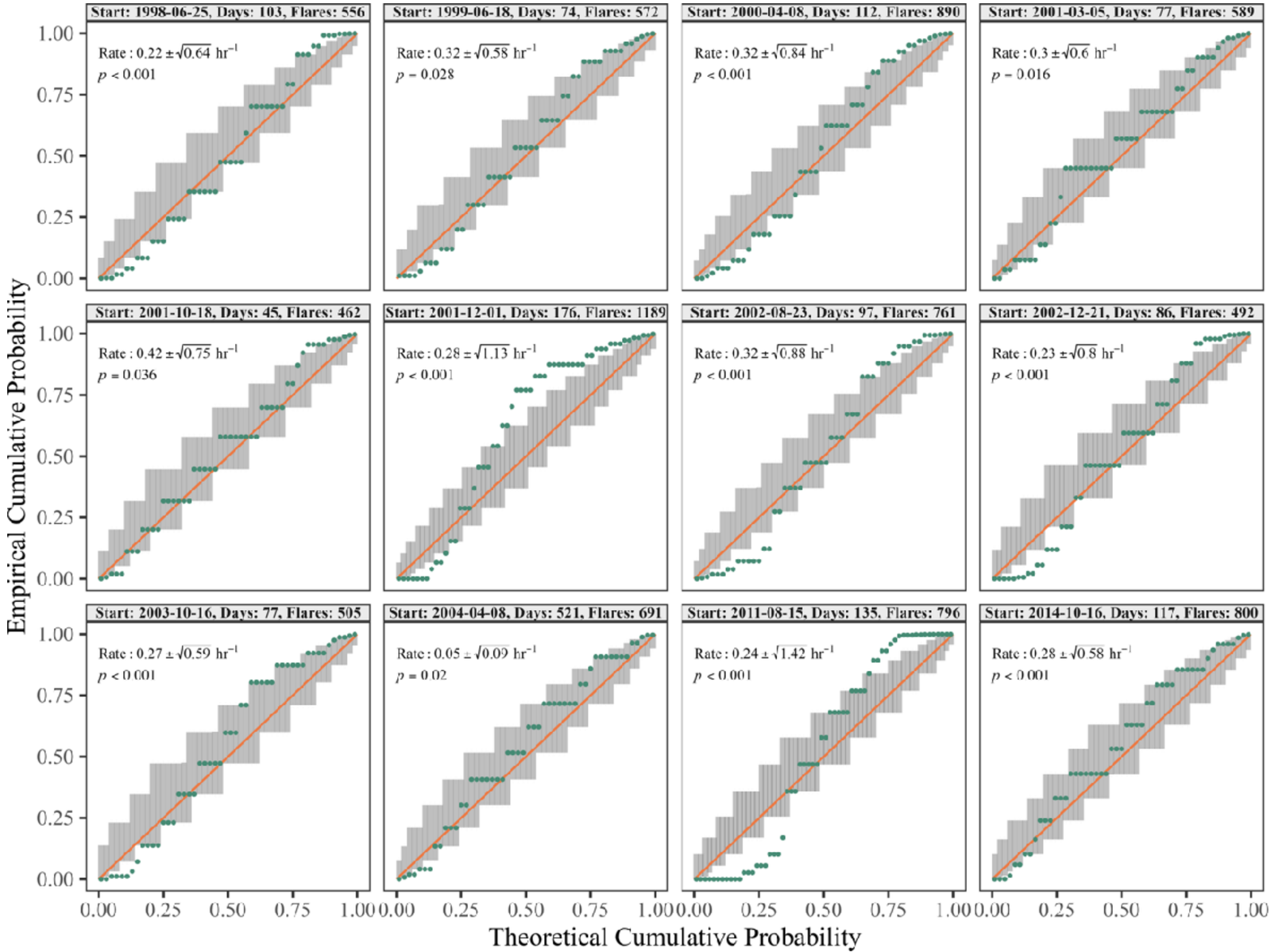
Result 1:

Aggregated flare count distributions are overdispersed

Overdispersion in Aggregate Count Distributions

- **Approach:**
  - Identify 342 distinct rate-constant time blocks using Bayesian Blocking method
  - Fit Poisson distribution to each block
  - Evaluate goodness-of-fit and overdispersion
- **Poisson unreasonable fit for ~20% of time blocks, of which 87% are overdispersed**

Min n	# Blocks	Non-Poisson	Overdispersed
30	203	42 (21%)	34 (81%)
50	153	31 (20%)	27 (87%)



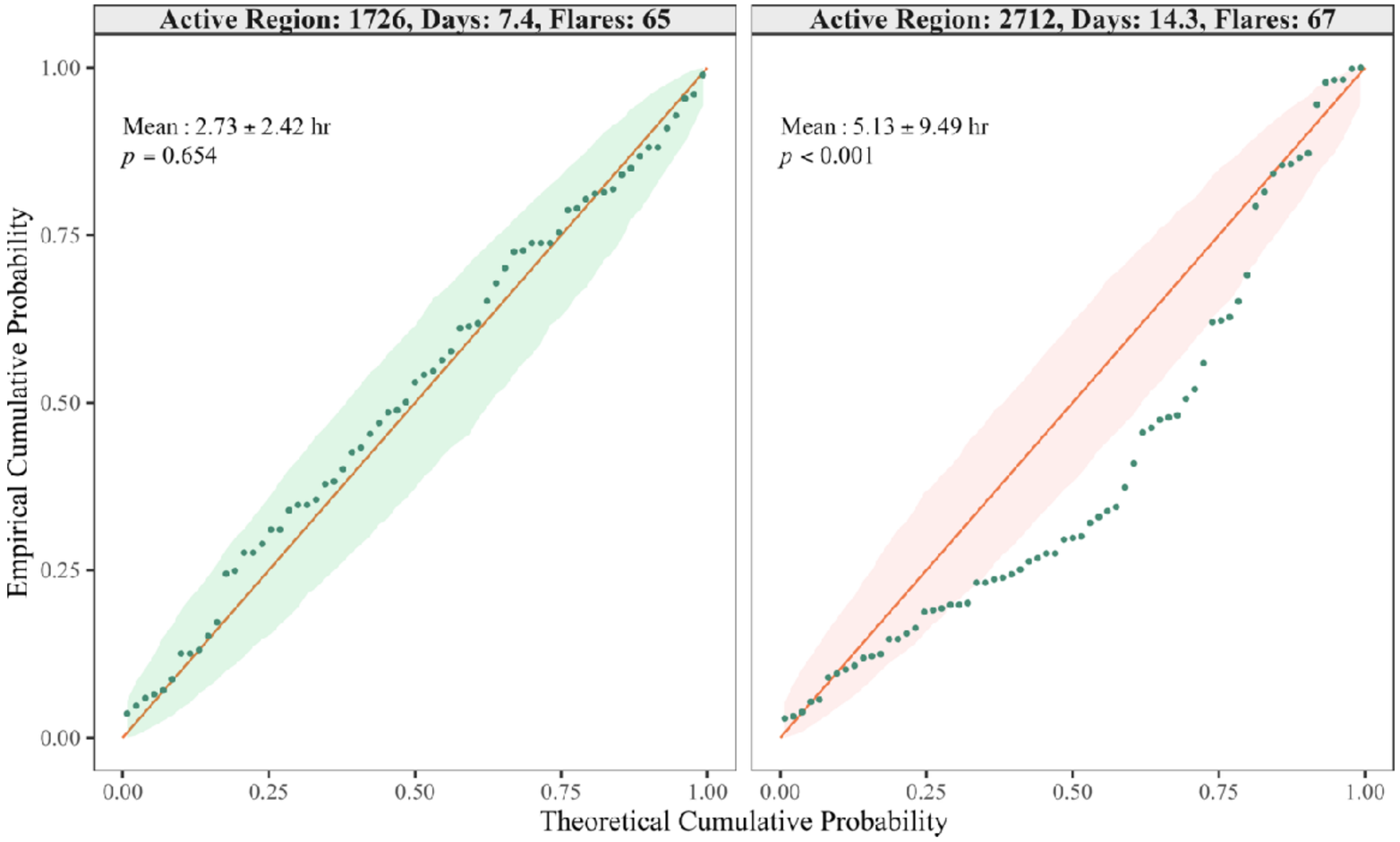
Result 2:

Flare waiting times within individual active regions depart from the exponential distribution and are overdispersed

Non-Exponential Waiting Times within Active Regions

- **Approach:**
  - Construct waiting time distributions for flares within each individual active region
  - Fit exponential distribution to waiting times
  - Evaluate goodness of fit and overdispersion
- **About 50% of active regions have non-exponential waiting time distributions, many of which are overdispersed**

Minimum n	# Active Regions	% Not Exponential
50	52	27 (52%)
30	144	71 (49%)

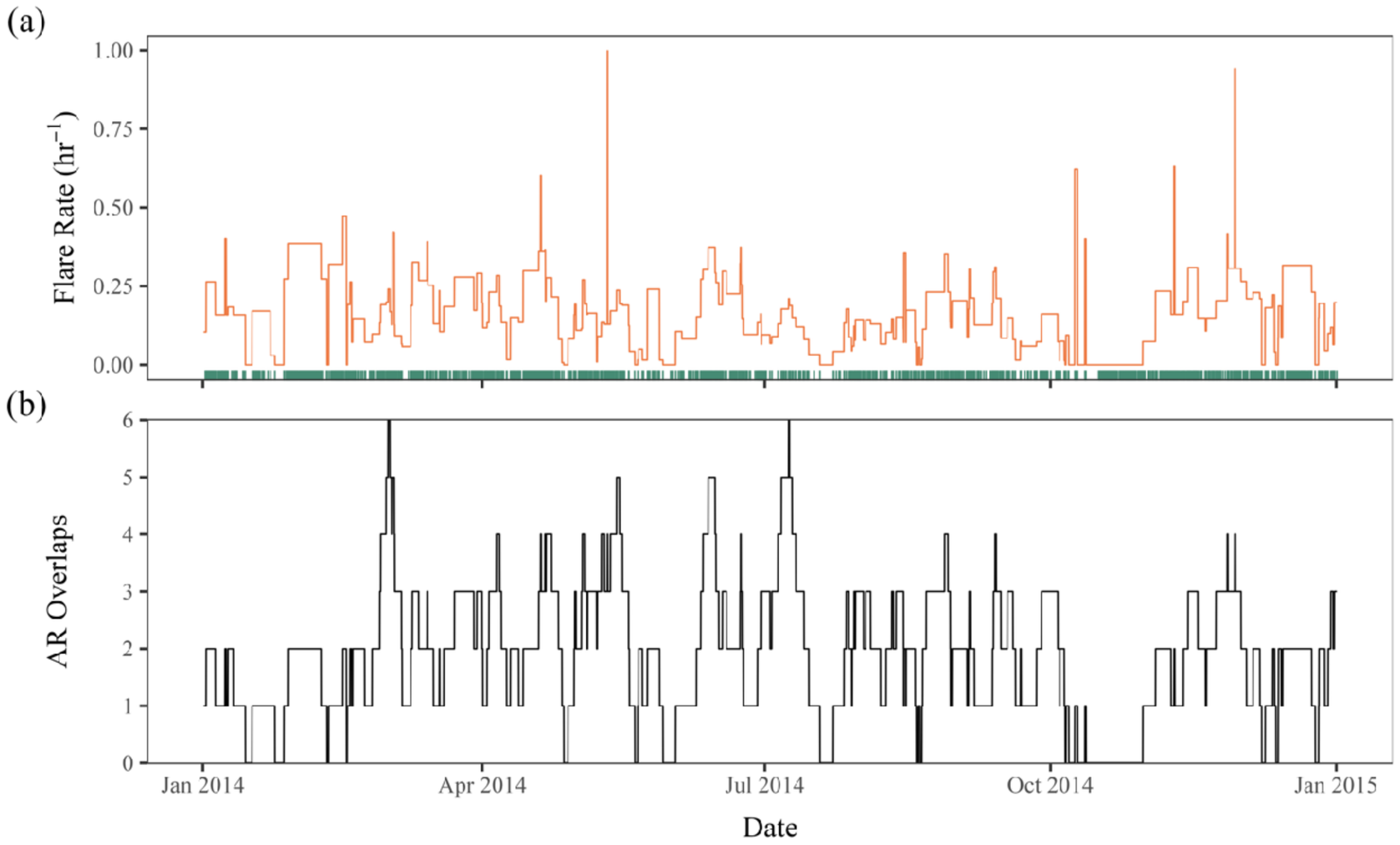


Result 3:

The observed counting process—a mixture of temporally overlapping active regions—results in overdispersion

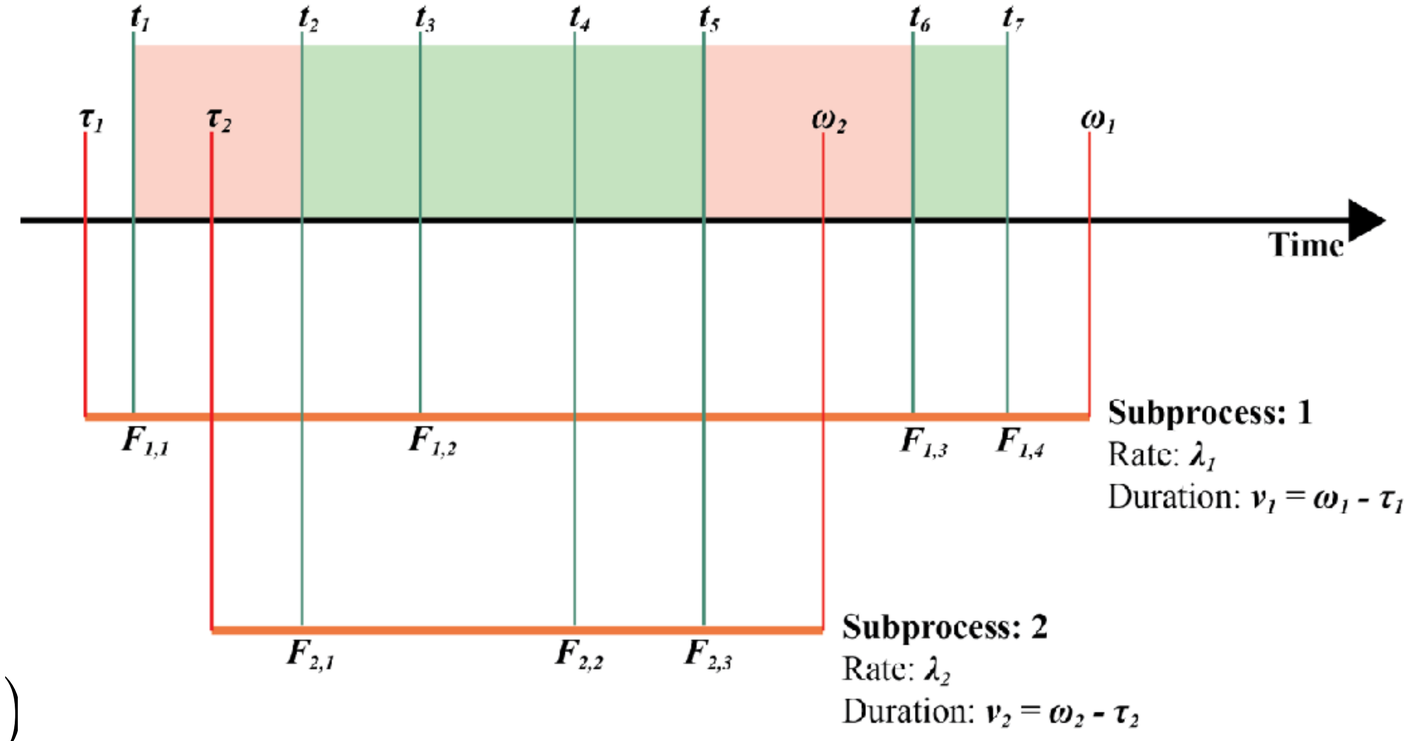
What Contributes to Overdispersion?

- We observe non-exponential waiting times and overdispersed counts, even within active regions. Why?
  - Non-stationary flare rate due to solar cycle
  - Sympathetic flaring within active regions (non-independence of events)?
  - **Unresolved, latent subprocesses from temporally overlapping active regions**



Overlapping Active Regions Model

- We model the observed event count process $N(t)$ as the sum of K **Poisson subprocesses** (K active regions), each active only on an interval $[\tau_k, \tau_k + v_k]$
- For each $k = 1,\ldots, K$ subprocess, we have:
  - Onset time: τ_k
  - Duration: $v_k = \omega_k - \tau_k$
  - Rate: λ_k
- For each subprocess: $N_k(t) \sim \text{Pois}\big(\lambda_k (t - \tau_k)\big)$



Overlapping Model

- The observable process is described by:

$$N(t) = \sum_{k=1}^{K} \left[N_k(t) \cdot \mathbb{I}(\tau_k < t \leq \tau_k + v_k) + N_k(\tau_k + v_k) \cdot \mathbb{I}(t > \tau_k + v_k) \right]$$

- The indicator functions ensure each subprocesses contributes only during its active period, and is fixed after

- We could fix the latent parameters and number of subprocesses, or set priors:

  - $K \sim p_K(\theta)$

  - $(\tau_k, v_k, \lambda_k) \sim f(\tau, v, \lambda)$

- For fixed K, fixed $\lambda_k = \lambda$ for all k, and disjoint support with no gaps, then $N(t)$ reduces to a stationary Poisson process with rate λ

Characteristic Function

- Use the characteristic function to study how this model departs from Poisson in various settings
- Generally, the characteristic function is:

$$\varphi_X(u) \;=\; \mathbb{E}\left[e^{\,i\,u\,X}\right] \;=\; \int_{-\infty}^{\infty} e^{\,i\,u\,x}\,dF_X(x), \quad u \in \mathbb{R}$$

where moments can be found via the Taylor series expansion

- Characteristic function for the overlapping model, given $(\tau_k, v_k) \sim f(\tau, v)$:

$$\varphi_{N(t)}(u) = \sum_{K=1}^{\infty} p_K(\theta) \left[\int_0^t \int_0^{t-\tau} e^{\lambda\,(t-\tau)\,(e^{iu}-1)} f(\tau, v)\,dv\,d\tau \;+\; \int_0^{\infty} \int_0^{t} e^{\lambda\,v\,(e^{iu}-1)} f(\tau, v)\,d\tau\,dv \right]^K.$$

Mean and Variance Derivation

- Setting:
  - $K = 2$ overlapping Poisson subprocesses, each of rate $\lambda > 0$
  - Fixed start times, with lifetimes $\sim exp(a)$ (mean duration $\frac{1}{a}$)
- First Moment:

$$\mathbb{E}[N(t)] = -i\,\varphi'_{N(t)}(0) = \frac{2\,\lambda}{a}\left(1 - e^{-at}\right)$$

- Second Moment:

$$\mathbb{E}[N(t)^2] = 2\left(\frac{e^{-at}\,\lambda}{a}\right)^2 + \frac{2}{a^2}\left[a\,e^{-at}\,\lambda\,(e^{-at} - 1) + 2\,e^{-at}\,\lambda^2\,(at + 1)\right]$$

Results: Mean and Variance

- Mean:

$$\mathbb{E}[N(t)] = -i\,\varphi'_{N(t)}(0) = \frac{2\,\lambda}{a}\left(1 - e^{-at}\right)$$

- Variance:

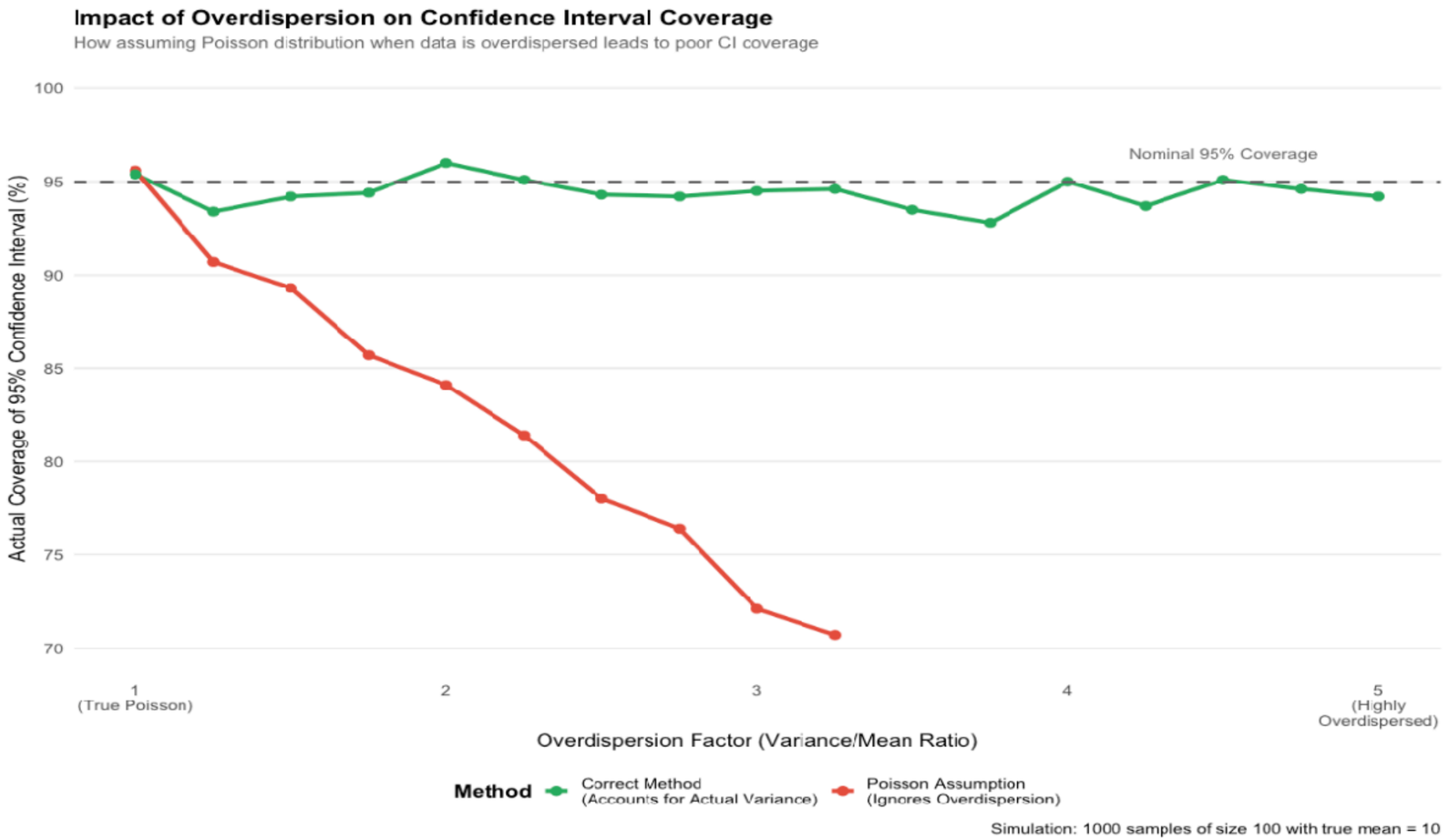
$$\text{Var}[N(t)] = \mathbb{E}[N(t)]\sigma(\lambda, a, t) = \frac{2\lambda}{a}(1 - e^{-at})\sigma(\lambda, a, t)$$

- Where $\sigma(\lambda, a, t) = \frac{\lambda}{a}(1 + e^{-at}) + \frac{1}{1 - e^{-at}}(1 - (1 + 2\lambda t)e^{-at})$ is the **overdispersion factor** and always ≥ 1

Moving Forward

Implications for flare studies

- Many flare studies derive uncertainties based on Poisson assumptions for counts[1]
  - In the presence of overdispersion, this assumption leads to systematic errors and underestimated uncertainties
  - Used to constrain physical processes as preparatory analysis for flare forecasting, so understanding overlapping process important[2]
- **Takeaways:**
  - **1. Check your assumptions!**
  - **2. Consider alternative models to account for overdispersion (e.g., negative binomial)**

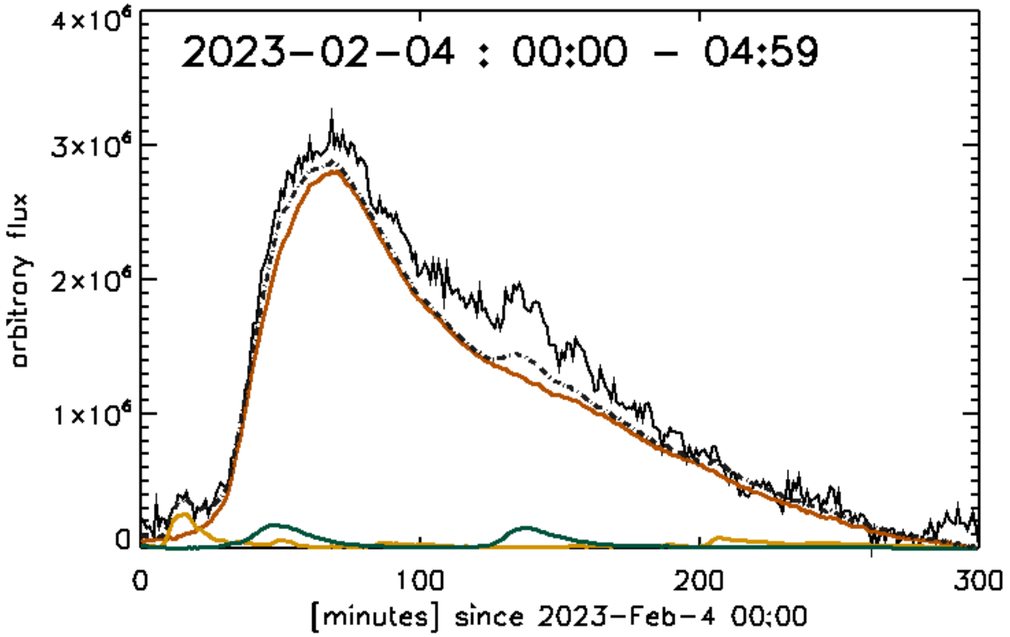


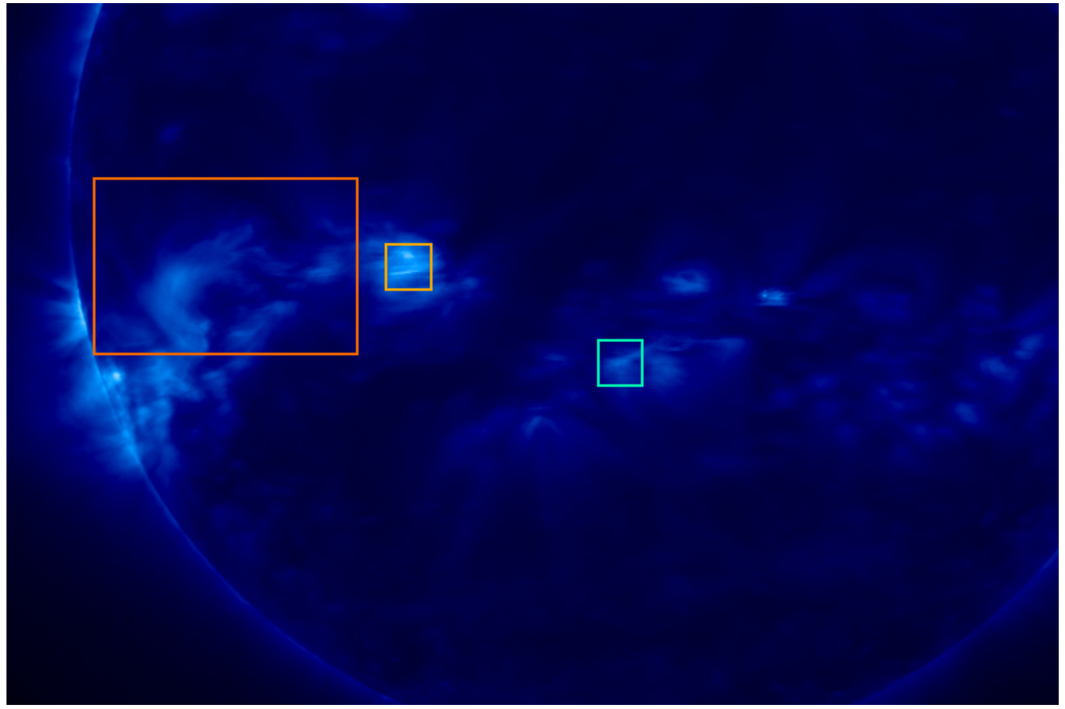
[1]Burton, K. et al. (2025). The Proxima Centauri Campaign - First Constraints on Millimeter Flare Rates from Alma.

[2]Biasiotti, L. & Ivanovski, S. L. (2025). Statistical Analysis of Solar Flare Properties from 1975 to 2017.

Limitations of Analysis

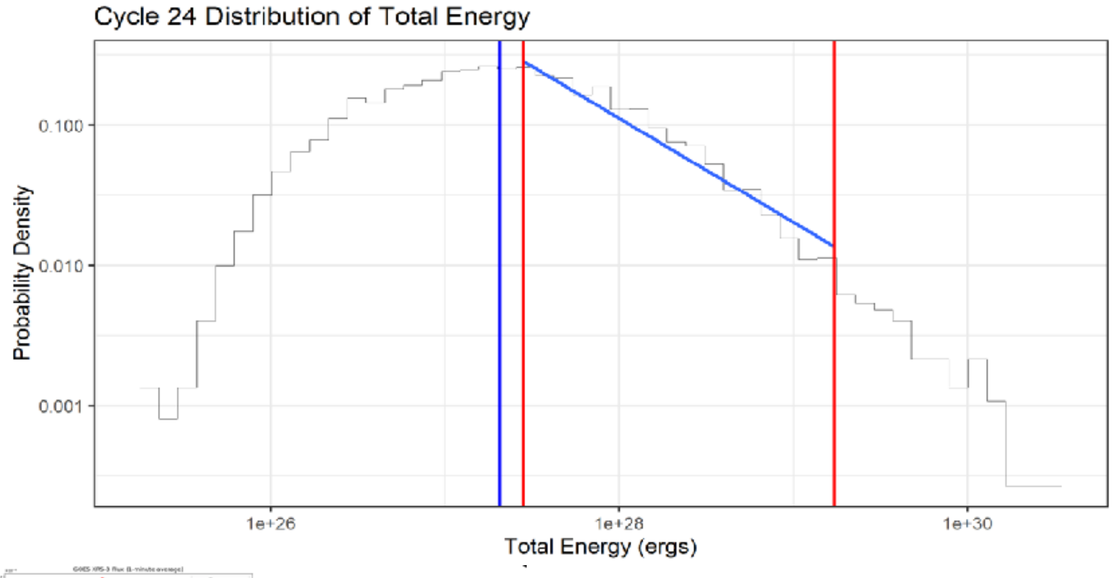
- Many low-energy flares missing in catalog
  - Difficult to identify temporally overlapping flares in full-disk light curve
  - Increased background flux makes weaker flares more likely to be undetected
  - Relatively few A and B class flares due to sensitivity and detection limitations

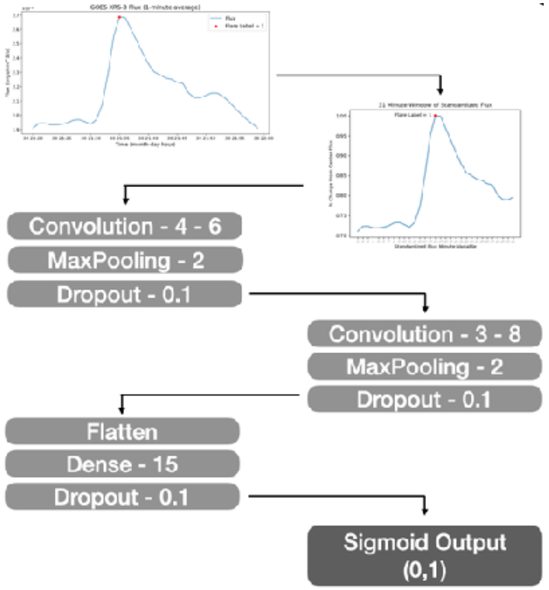


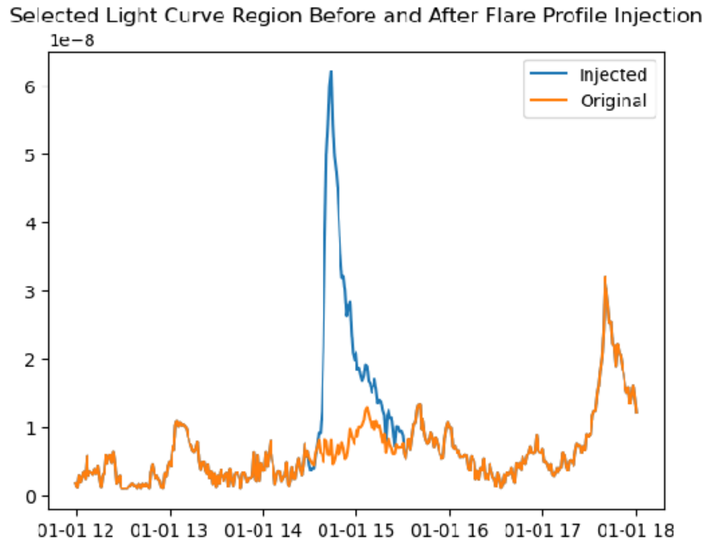


Future Work

- Generate more comprehensive flare catalog using wavelets and convolutional neural networks[1]
- Fit power-laws to distribution of flare energy distribution with new catalog
- Connect results to constrain physical processes and improve flare forecasting







[1] Ingram et al., (2023). Machine Learning Methods to Detect More Solar Flares. Parker Heliophysics Scholars 6th Meeting.

Acknowledgements

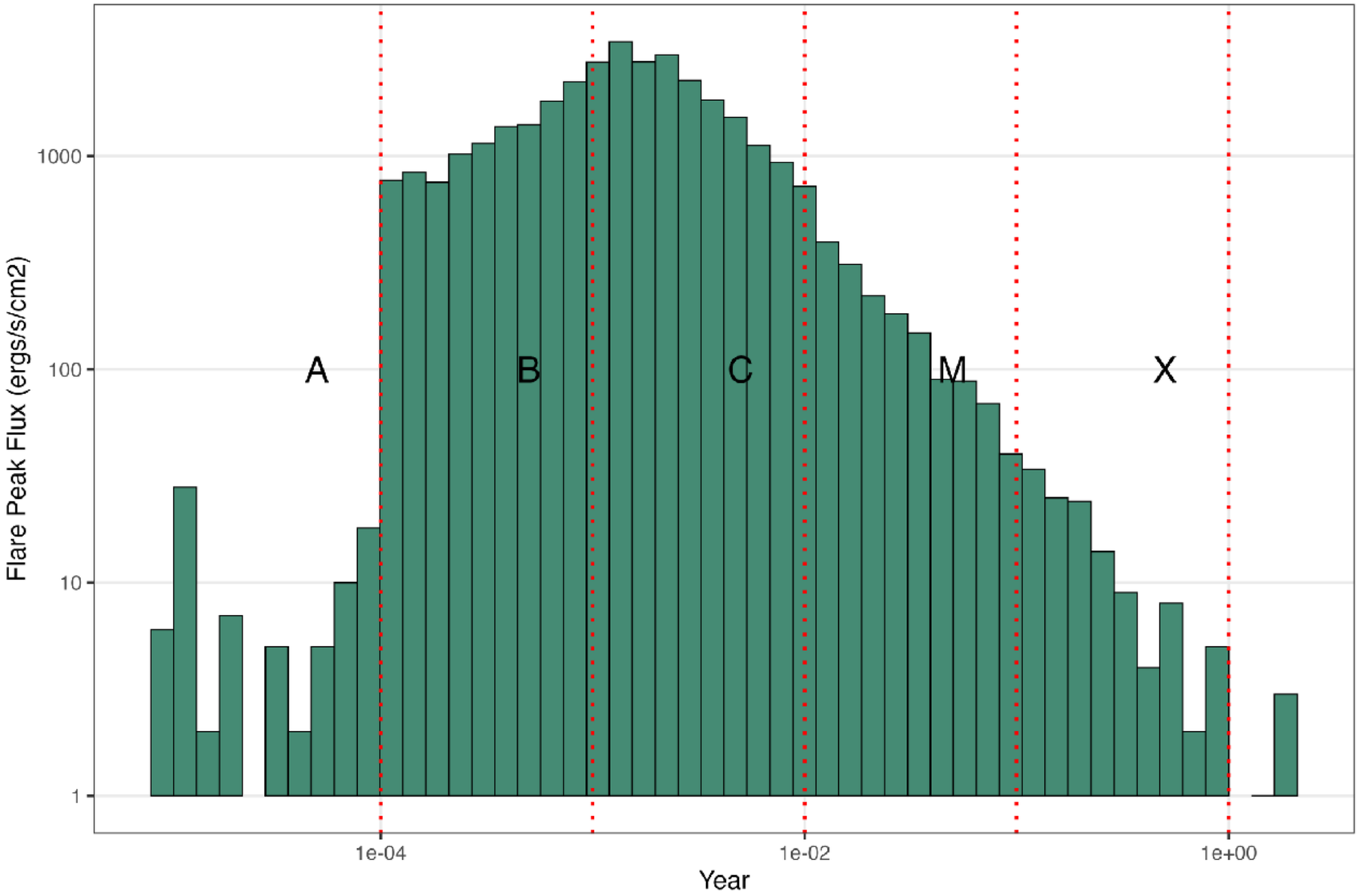
- This material is based upon work supported by the National Science Foundation Graduate Research Fellowship Program under Grant No DGE2140739. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

Summary

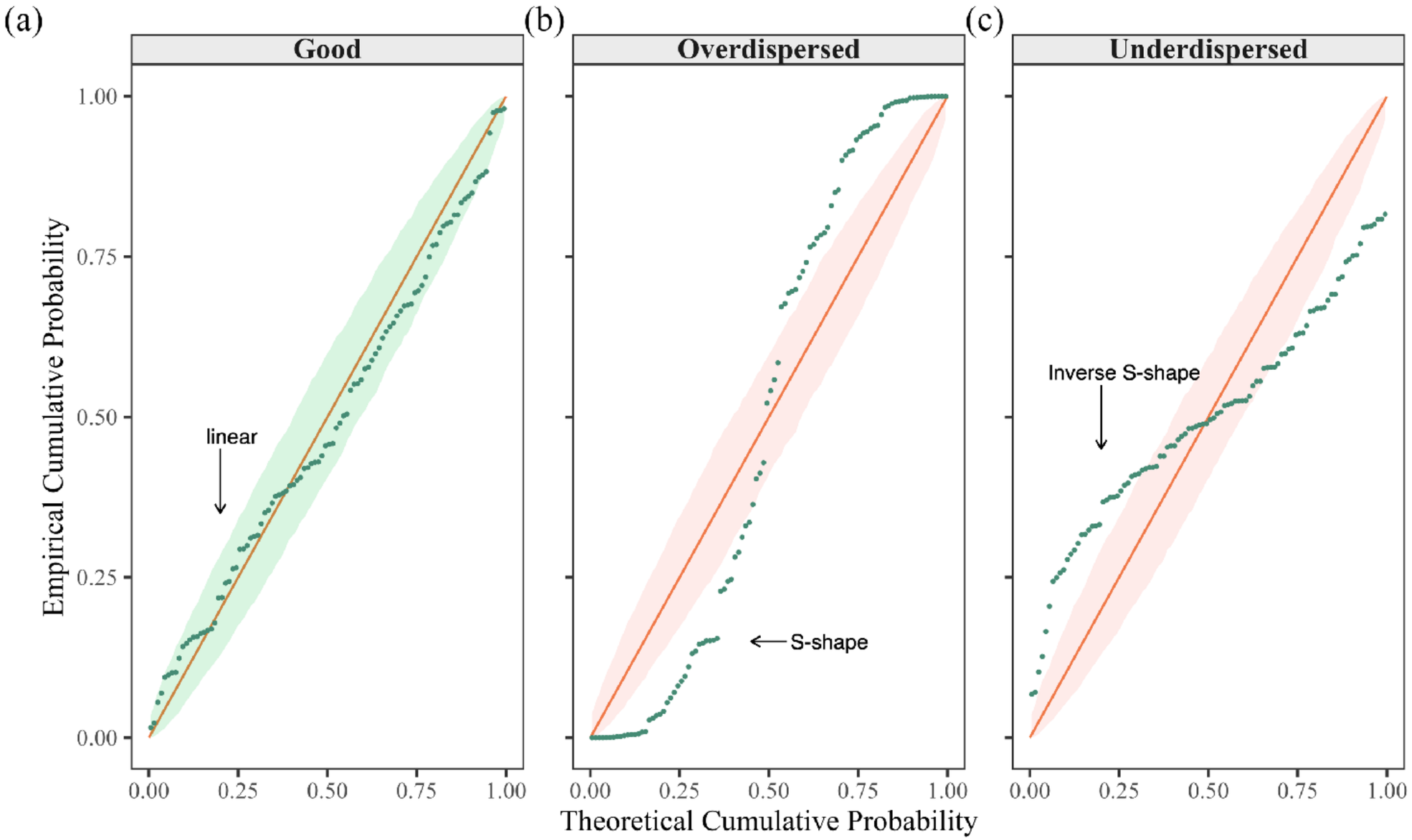
- **Takeaway:** *Be careful when making a Poisson assumption for flares; you may underestimate your uncertainties!*
- **Result 1:** Aggregate flare count distributions are overdispersed
- **Result 2:** Flare waiting times within individual active regions depart from the exponential distribution and are overdispersed
- **Result 3:** A statistical model for the observed counting process—a mixture of temporally overlapping active regions—results in overdispersion
- **Next:** Generate a more comprehensive flare catalog and fit power-law to flare energies

Supplementary Slides

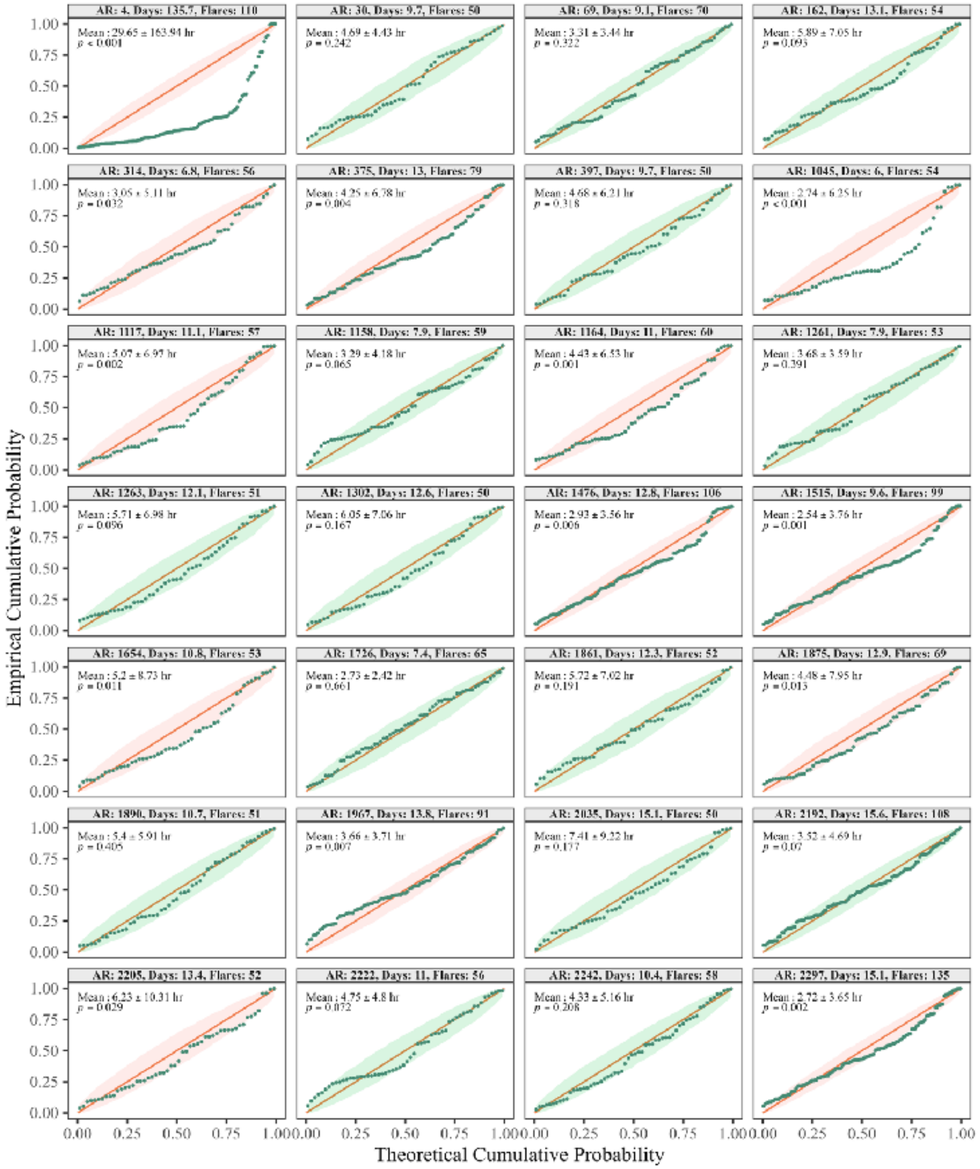
Missing low-energy/flux flares

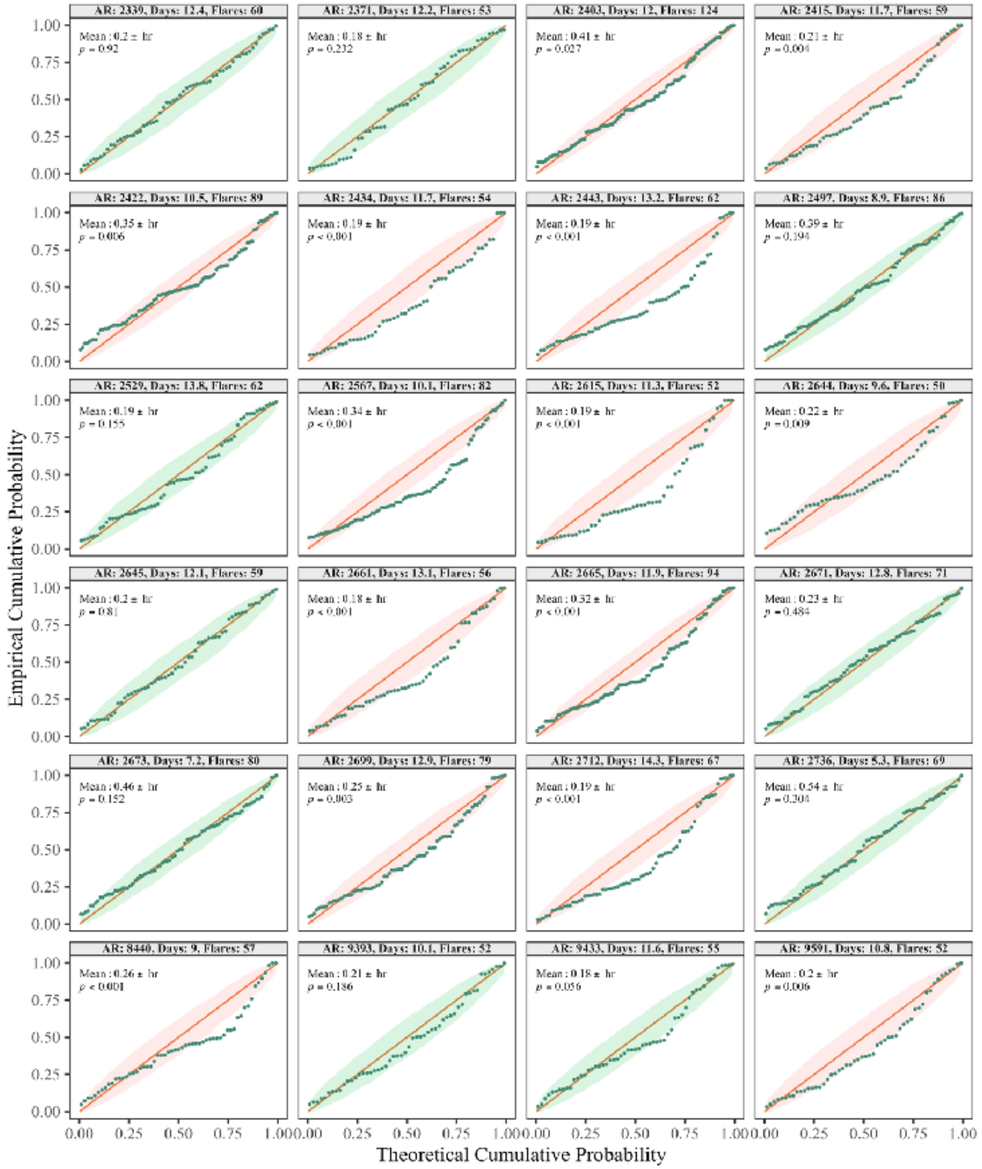


Interpreting Probability-Probability Plots

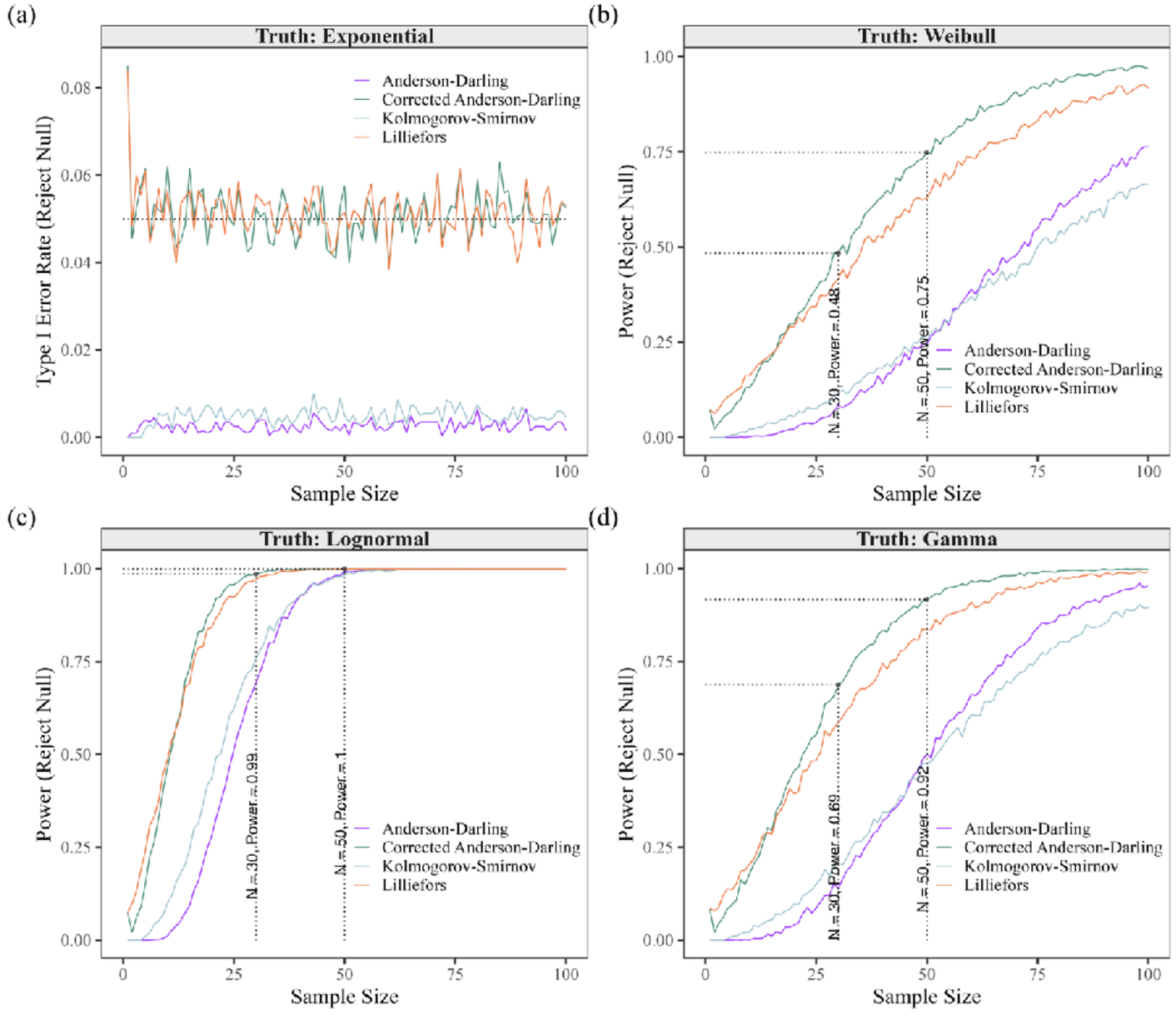


Active Region Exponential Fits

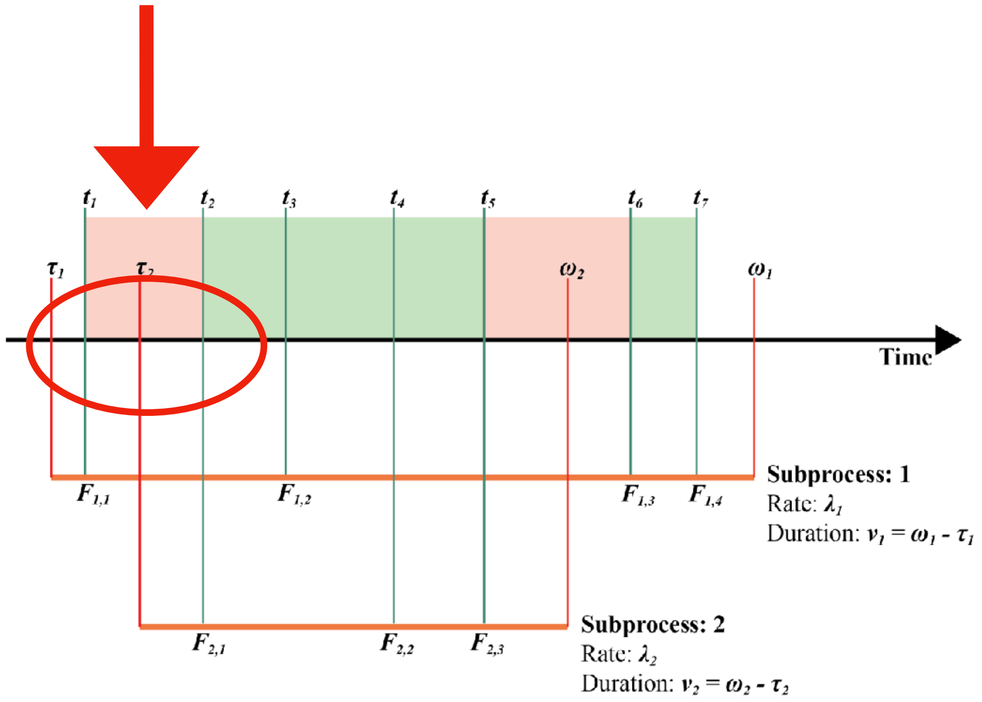


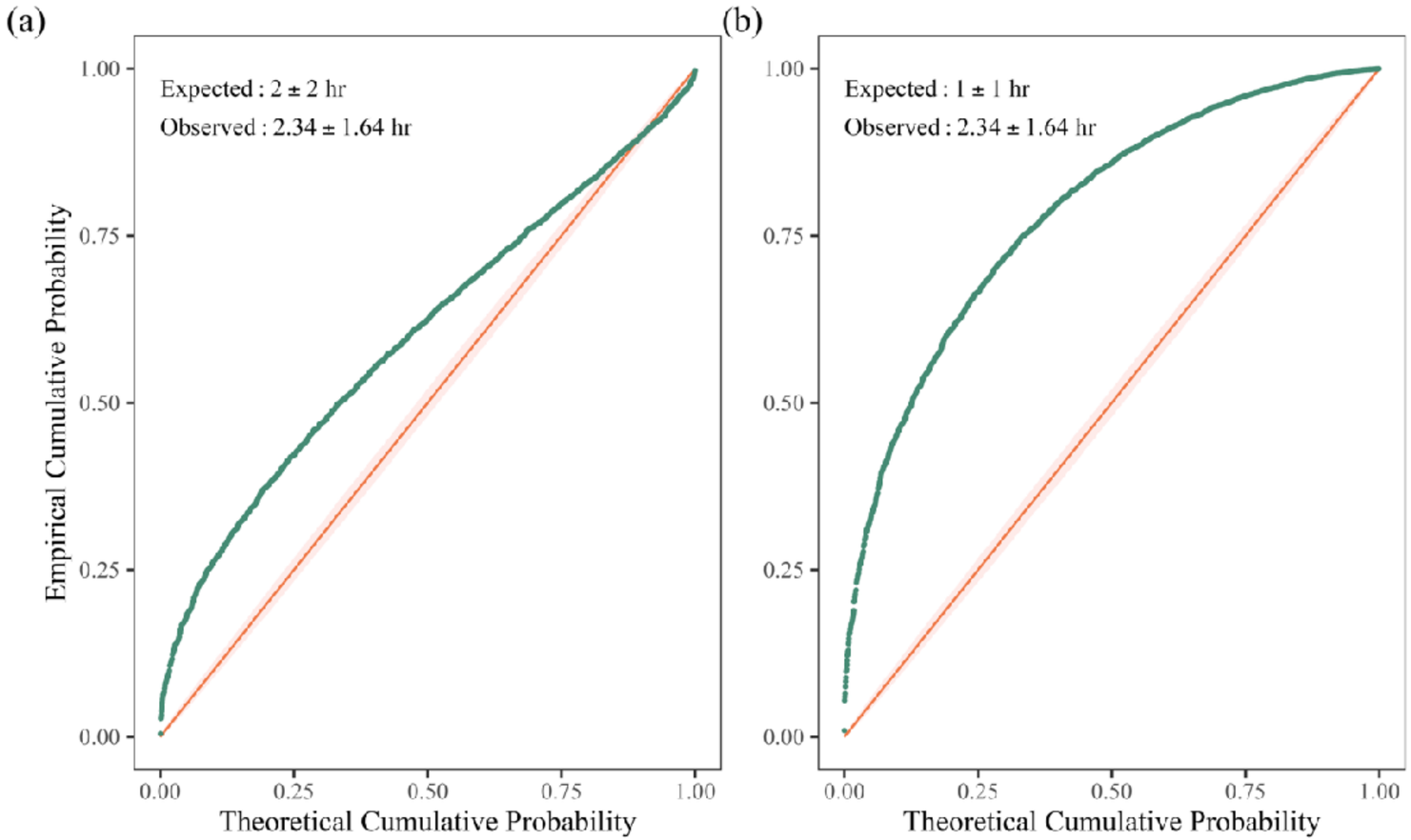


Power Analysis for Goodness of Fit Tests

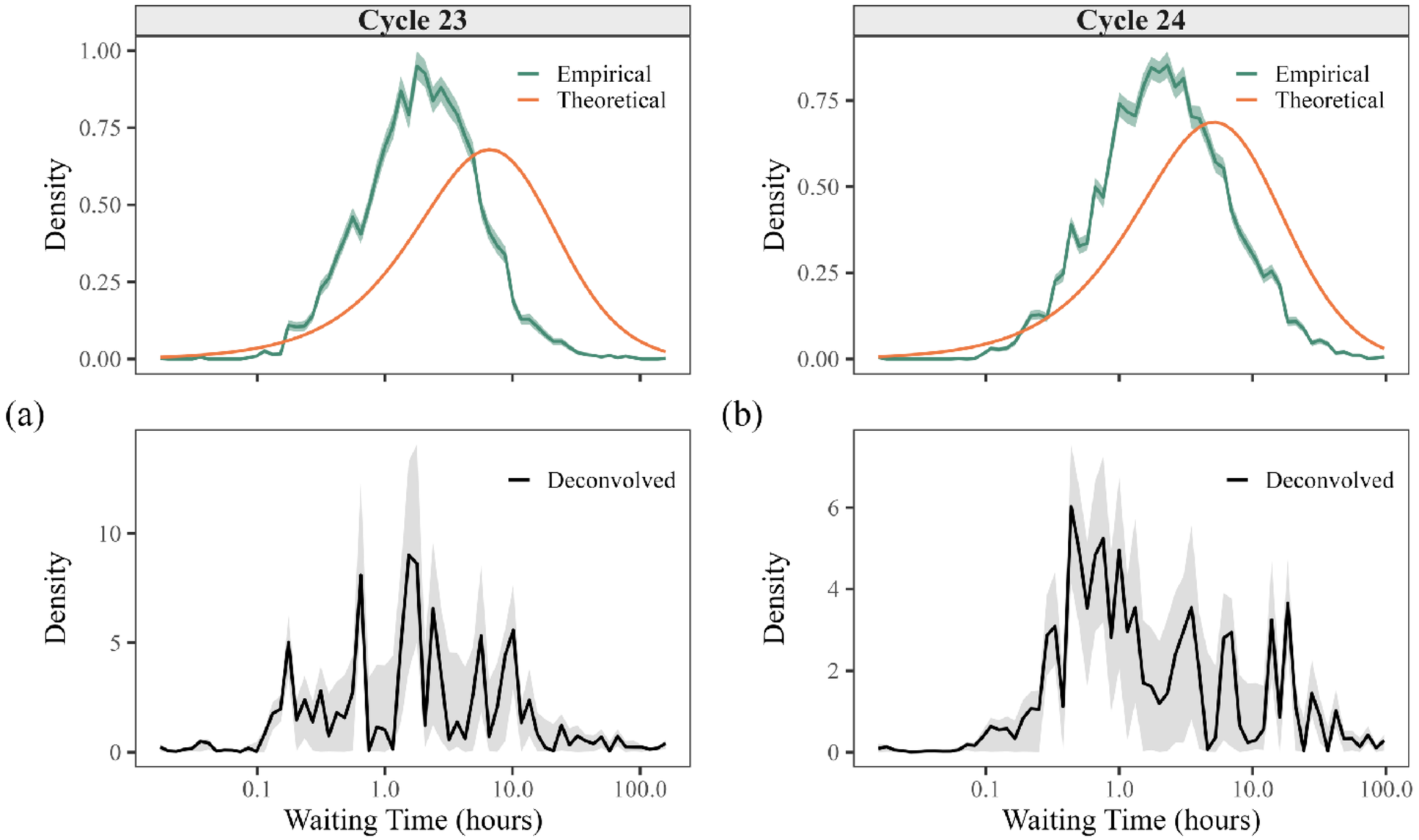


Overlapping Simulations at Point of Overlap

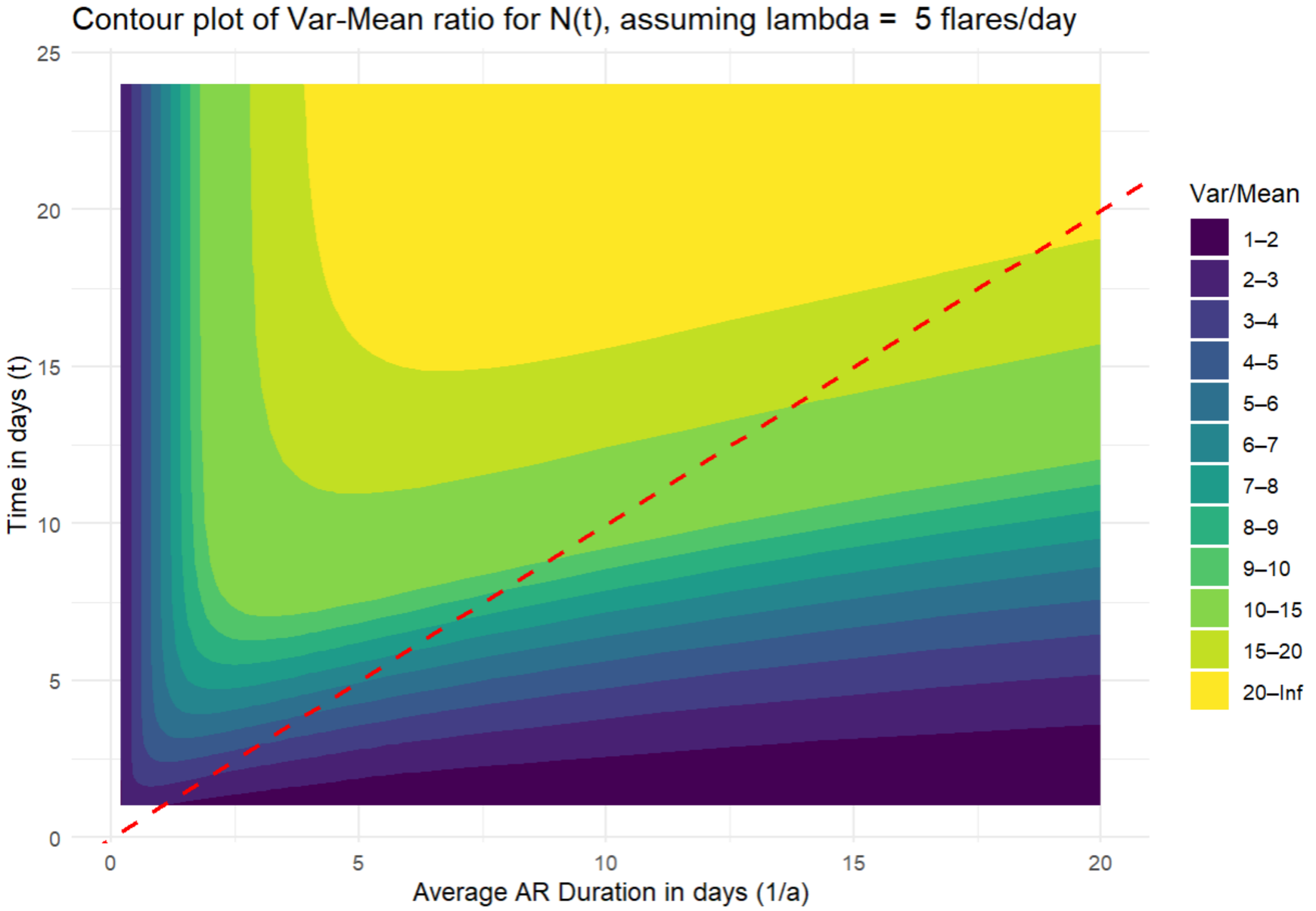




Deconvolution of Waiting Times

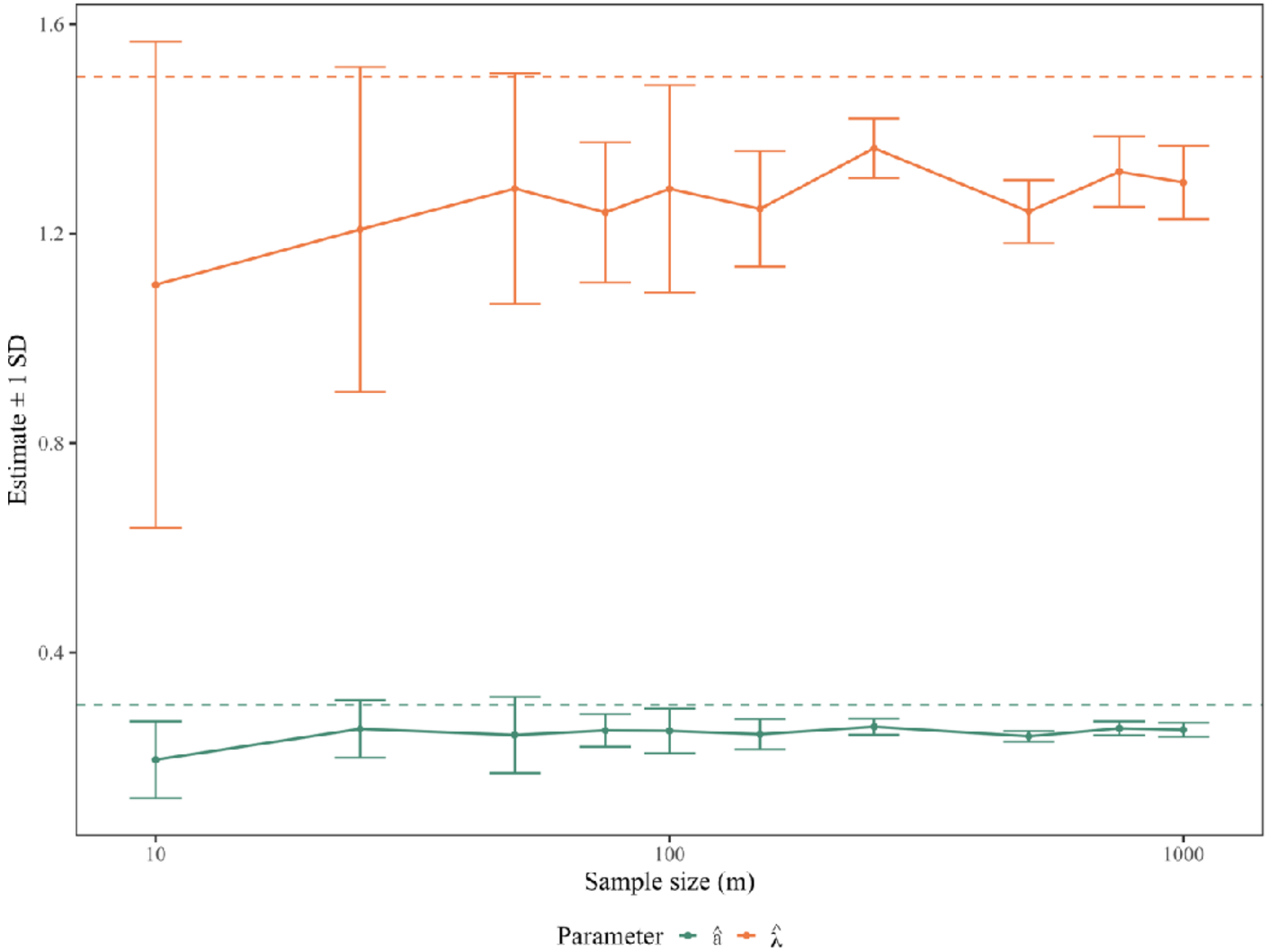


Overdispersion Contour Plot for Model



Estimation for Overlapping Model

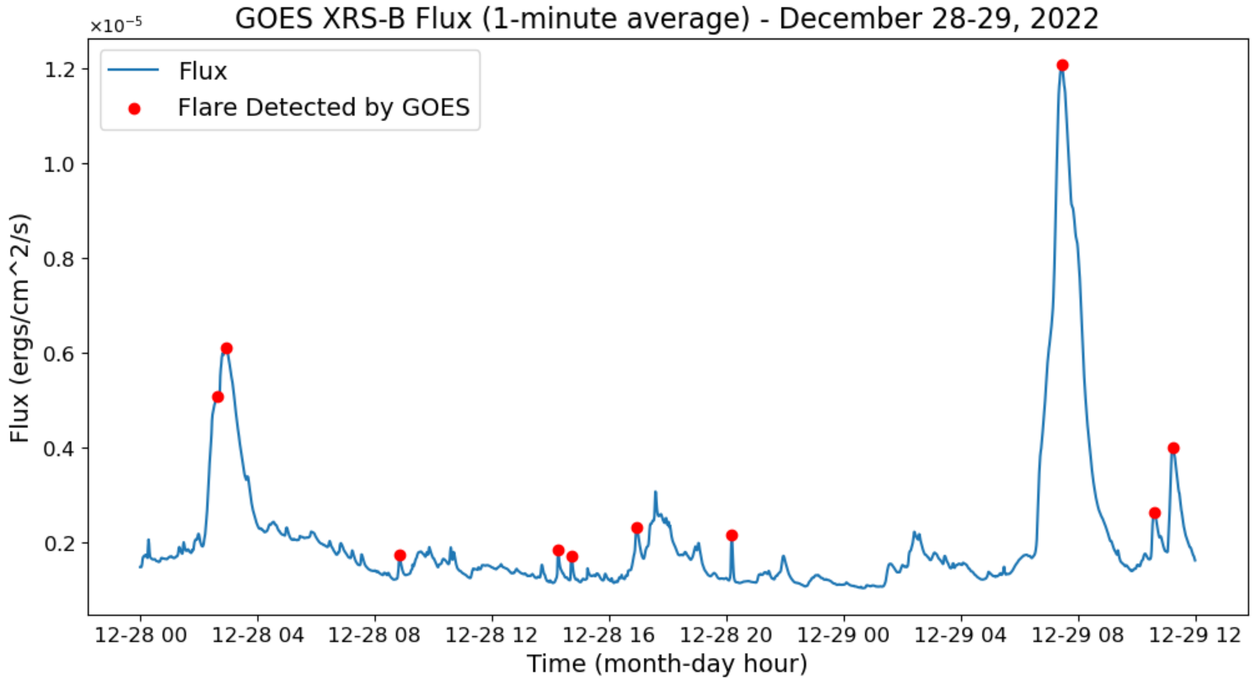
- Expectation-Maximization Algorithm



Flare Detection Using CNNs

GOES Solar Activity Tracking and Flare Detection

- GOES X-Ray Sensor (XRS) measures full-disk integrated flux in 0.5 - 4 Å (XRS-A) and 1-8 Å (XRS-B) passbands.
- Light curves are constructed from 1-minute average flux measurements for both passbands.
- Spikes in light curves correspond to flare event occurrence somewhere on the disc or limb.
- GOES flare catalog is constructed using a detection algorithm applied to the 1-8 Å passband.
    - Start time of flare is first minute of steep monotonic increase in flux in a sequence of 4 minutes.
    - Flare peak time is the minute of the maximum flux in the event time interval.
    - End time of flare is the minute of median of 3 successive background-subtracted intensities falls to 1/2 the background-subtracted peak.

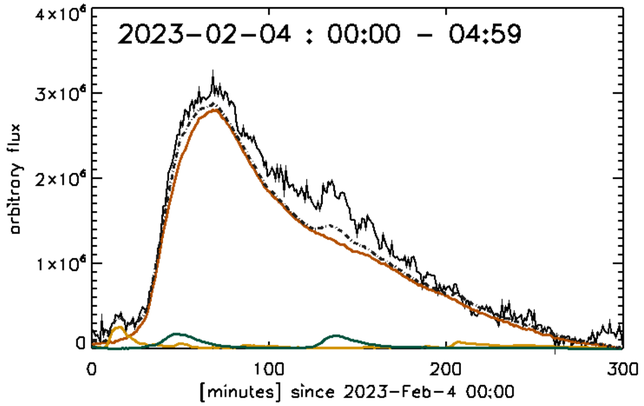


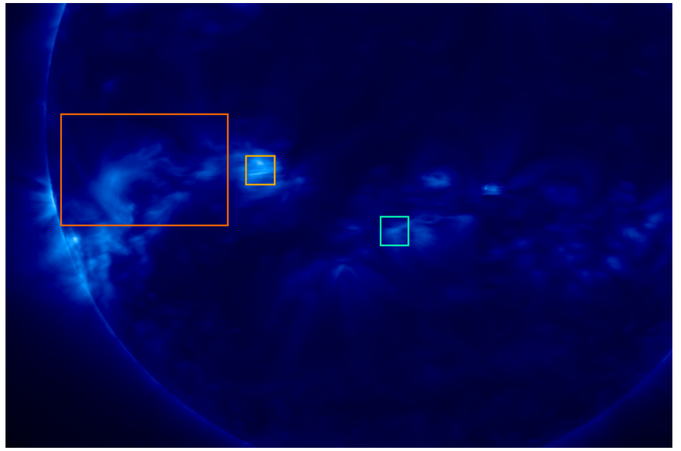
GOES detection algorithm described in-depth here: XRS User Guide

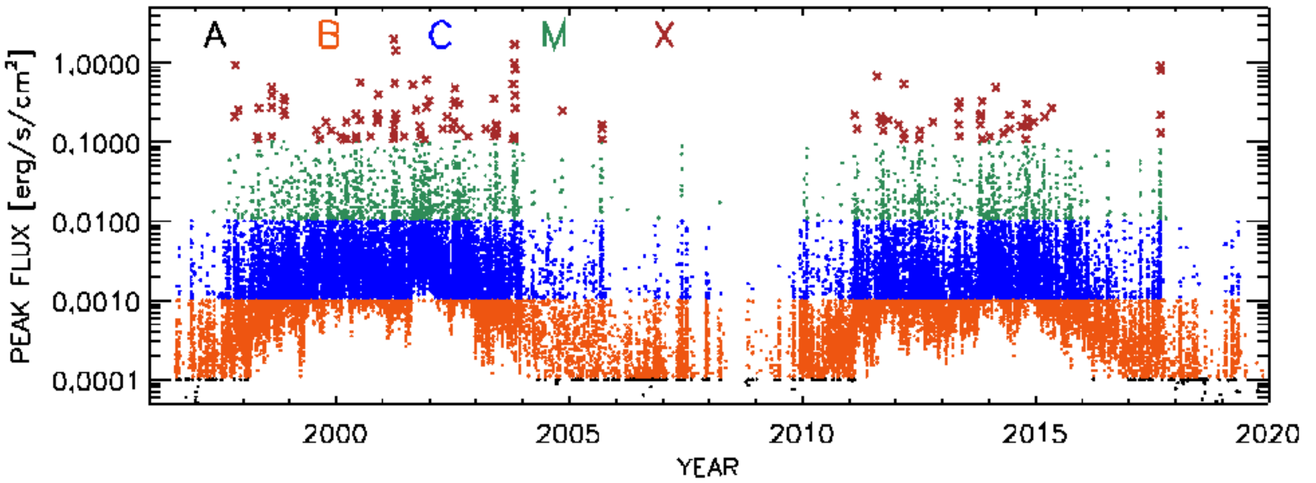
See data source: https://www.ngdc.noaa.gov/stp/satellite/goes-r.html

When are Flares Missed by GOES?

- Difficult to identify temporally overlapping flares in full-disk light curve.
- Increased background flux makes weaker flares more likely to be undetected.
- Relatively few A and B class flares due to sensitivity and detection limitations.

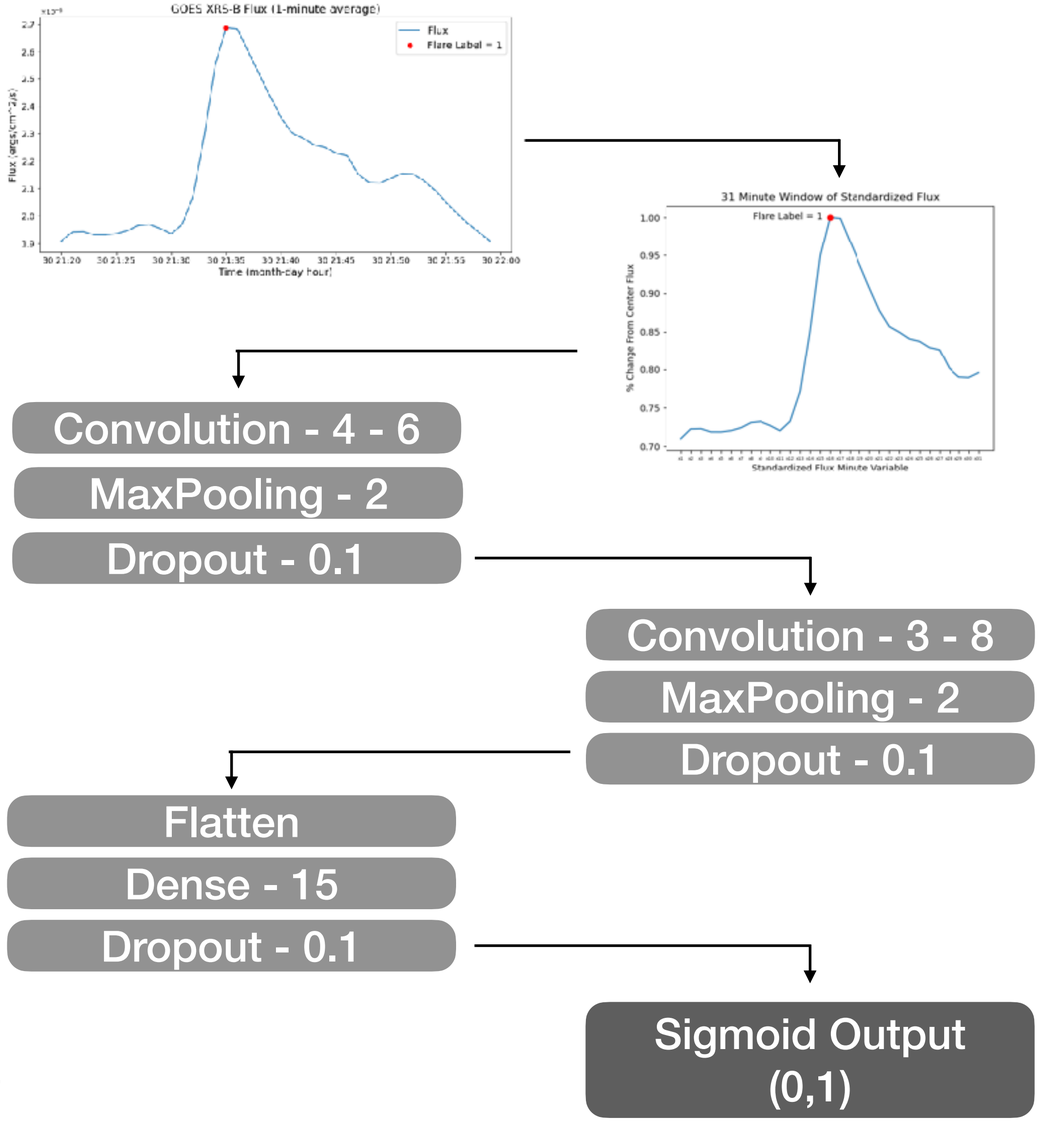






Convolutional Neural Networks

- Convolutional neural networks (CNNs) are used for anomaly detection in 1D time series data.
- To start, we follow the architecture from Feinstein et al. (2020)*, who used CNNs to detect stellar flares in TESS light curves.
- We train the CNN to identify solar flare peaks in ± 15 minute windows of standardized flux.
    - Standardized = % change from central flux in window.



*Feinstein et al. *Flare Statistics for Young Stars from a Convolutional Neural Network Analysis of TESS Data.* ApJ 160, 219, 2020.

Training the CNN

- Data consists of 156,727 31-minute windows labeled 1 for flare event peak at center, 0 for non-flare event.
  - 7,870 windows centered on GOES-detected flare profiles.
  - 148,857 windows of non-flare profiles determined by regions defined by overly sensitive wavelet method.
  - 80/10/10 train, validation, test data split.
  - Probability score threshold optimized with a validation set.

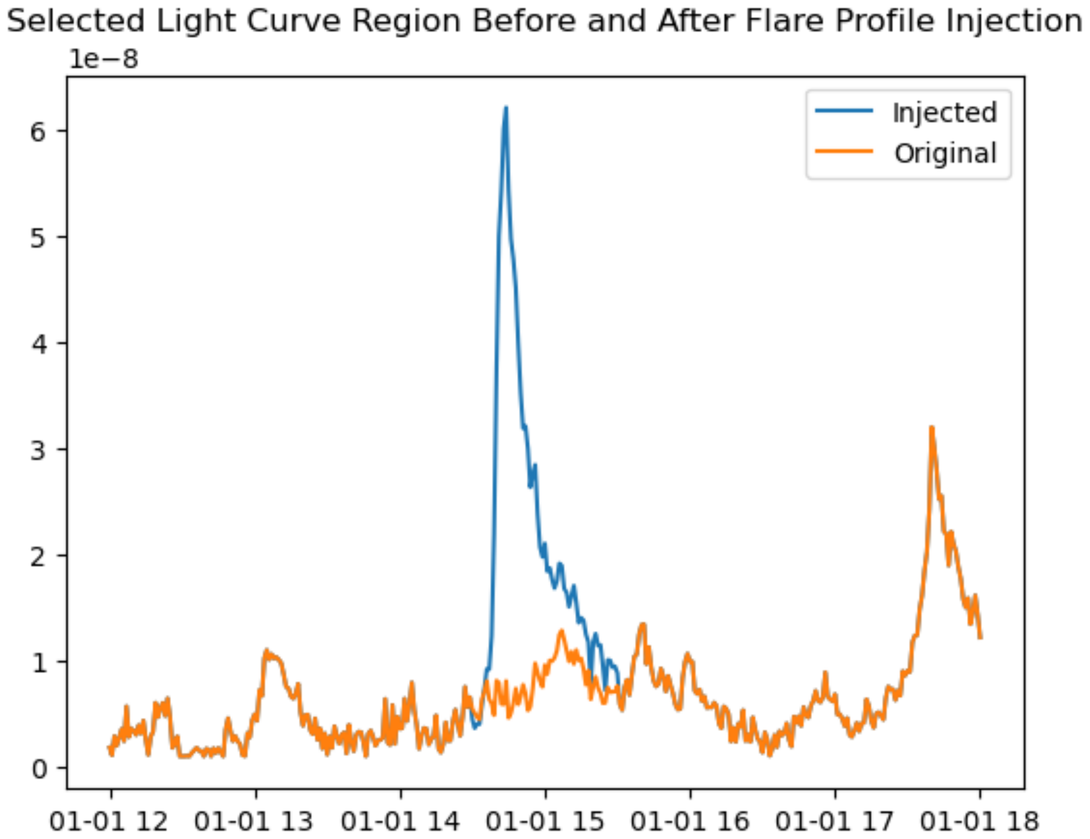
	Binary Accuracy	**Precision**	**Recall**
	$\frac{TP + TN}{N}$	$\frac{TP}{TP + FP}$	$\frac{TP}{TP + FN}$
Training Set (Threshold = 0.5)	99.31%	97.26%	88.68%
Test Set (Threshold = 0.05)	98.93%	84.82%	96.25%

TP = True Positive
TN = True Negative
N = Number of Predictions

FP = False Positive
FN = False Negative

How Do We Evaluate the CNN?

- Evaluate the CNN by generating synthetic flare and non-flare profiles.
- Select set of real flare profiles are injected into random regions of light curves to generate synthetic flare profiles.
- Synthetic non-flare profiles generated from adding A-class flat background to noise obtained from wavelet method.
- 7,550 synthetic flare profiles, 7,900 synthetic non-flare profiles.



CNN Performance on Synthetic Data

- CNN performs well on identifying synthetic flare profiles peaks.
- Next version of synthetic data needs to be more comprehensive to study where CNN succeeds and fails in a multitude of cases.

	Binary Accuracy	Precision	Recall
Evaluation Set (Threshold = 0.05)	96.79%	96.68%	96.75%

	Predicted Non-Flare	Predicted Flare	Total
True Non-Flare	7635	251	7886
True Flare	245	7303	7545
Total	7880	7554	15434