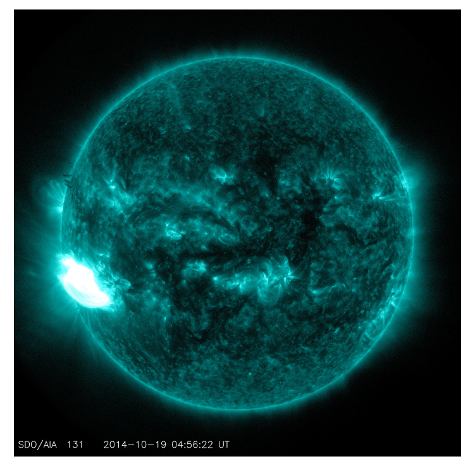
Interpreting LSTM prediction on Solar Flare Eruption with Time-series Clustering

Hu Sun[1], Ward Manchester[2], Zhenbang Jiao[1], Xiantong Wang[2], Yang Chen[1]

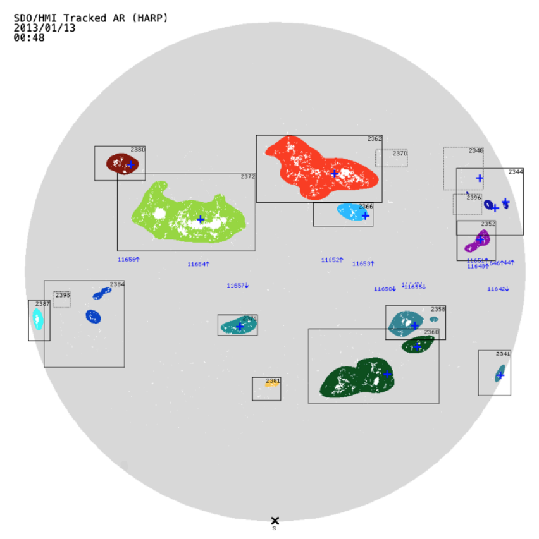
[1] Department of Statistics, University of Michigan, Ann Arbor [2] Department of Climate and Space Sciences and Engineering, University of Michigan, Ann Arbor

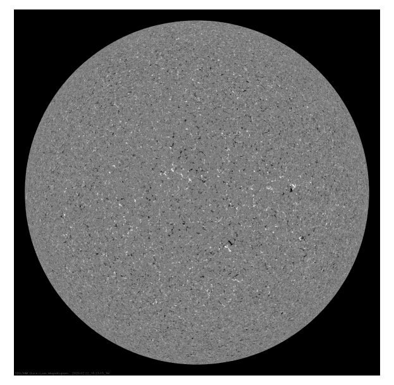
Introduction & Background



A solar flare is a type of eruptive activity, which occurs with a sudden increased brightness across the electromagnetic spectrum including radio waves and gamma-rays. The emission is observed in the extreme ultraviolet (EUV) with the morphology of loops in close proximity to a sunspot group. Strong solar flares may create disruptions in Earth's upper atmosphere and hamper signal transmission.

Solar flares are observed in **active regions** of the Sun (shown as rectangle boxes on the right, which represents the Helioseismic and Magnetic Image (HMI) Active Region Patches). For every active region, we have a list of solar flare records from Geostationary Operational Environmental Satellite (GOES). Each flare is categorized based on the soft X-ray intensity into **classes labeled by A, B, C, M and X**, where X flares are the strongest.





Our machine learning task is to classify strong solar flares (M and X class flares) against weak solar flares (B flares) based on Space Weather HMI Active Region Patches (SHARP) parameters hours before each flare.

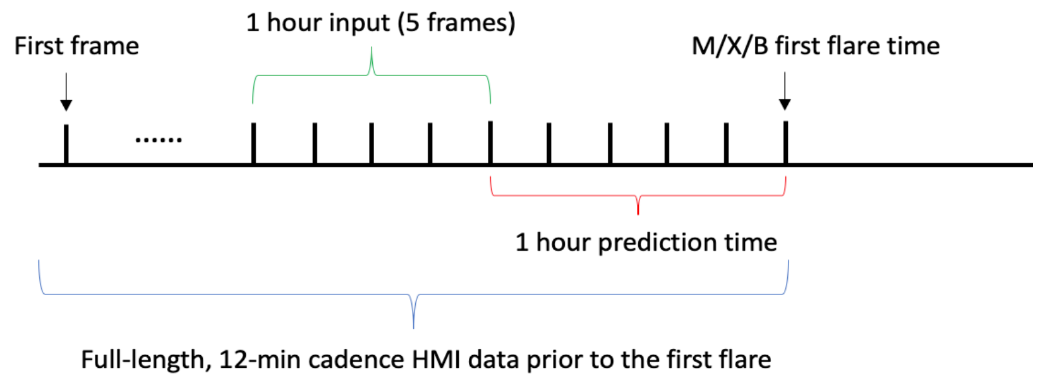
Data & Machine Learning Model

- Among all 10,000+ recorded flares during 2010-Dec to 2018-Jun coming from the 860 active regions, we are interested in the **first B/M/X flares.** In total, we have 97 strong flares and 305 weak flares coming from 369 active regions after discarding flares with >10% of missing frames.

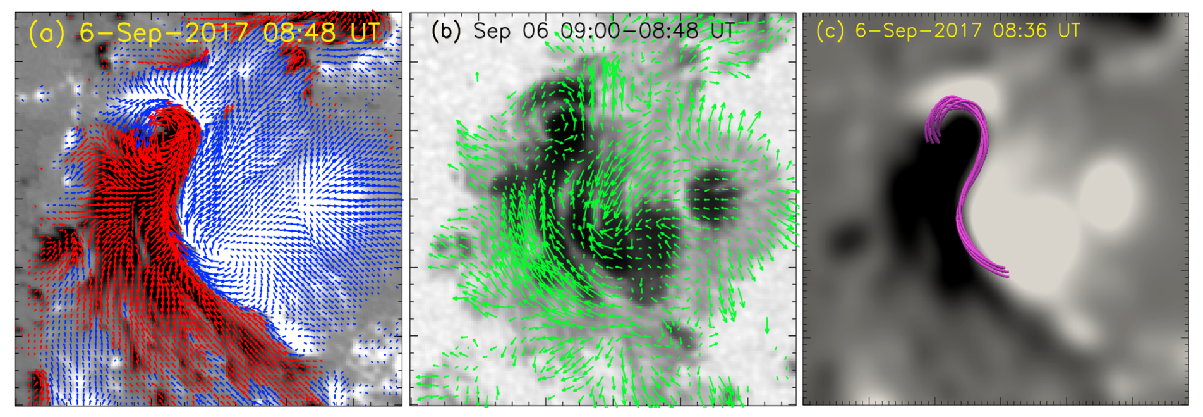
Class/Year	2010	2011	2012	2013	2014	2015	2016	2017	2018	Total
M/X	0	15	12	20	27	18	3	2	0	97
B	2	53	48	52	23	46	52	23	6	305

Table 1: Flare Data Count Summary across 2010-2018

- Our predictor data for each flare is collected from the 12-min cadence SHARP parameter data. 1 hour before the recorded flare time, we collect 5 frames of HMI images for the corresponding active region:

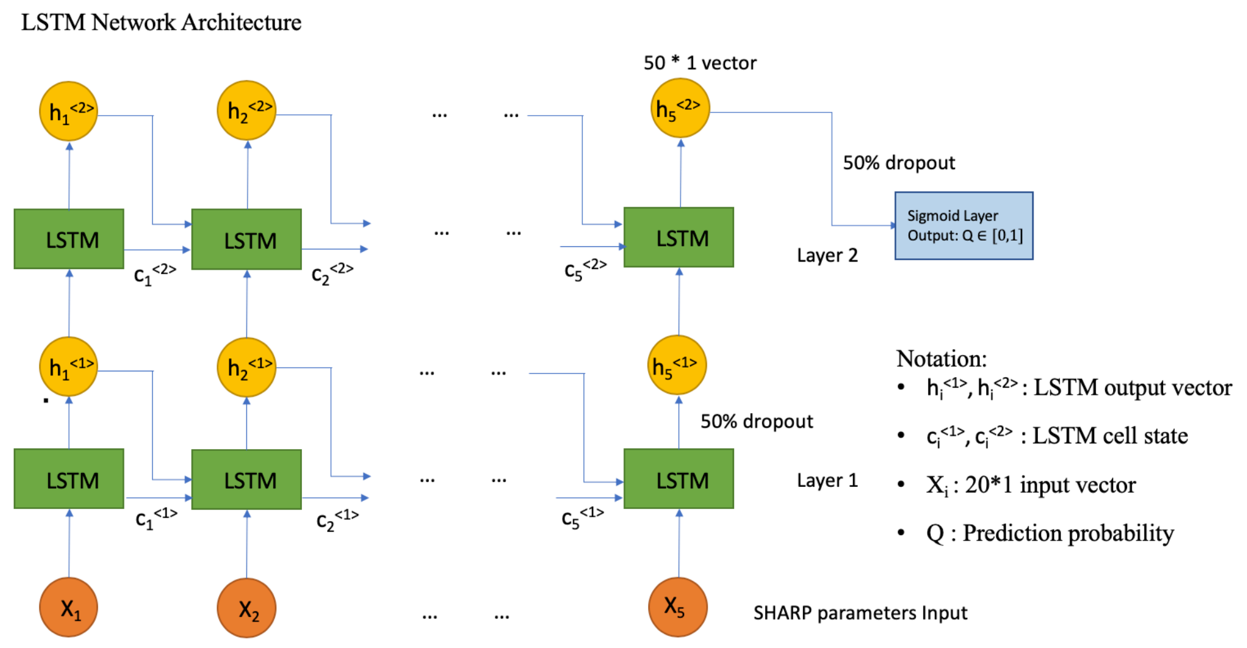


- For each frame of HMI image, we highlight the pixels along a special structure called polarity inversion line (PIL) which is considered to be the key region related to flare eruption. PIL is the boundary splitting positive and negative magnetic field:



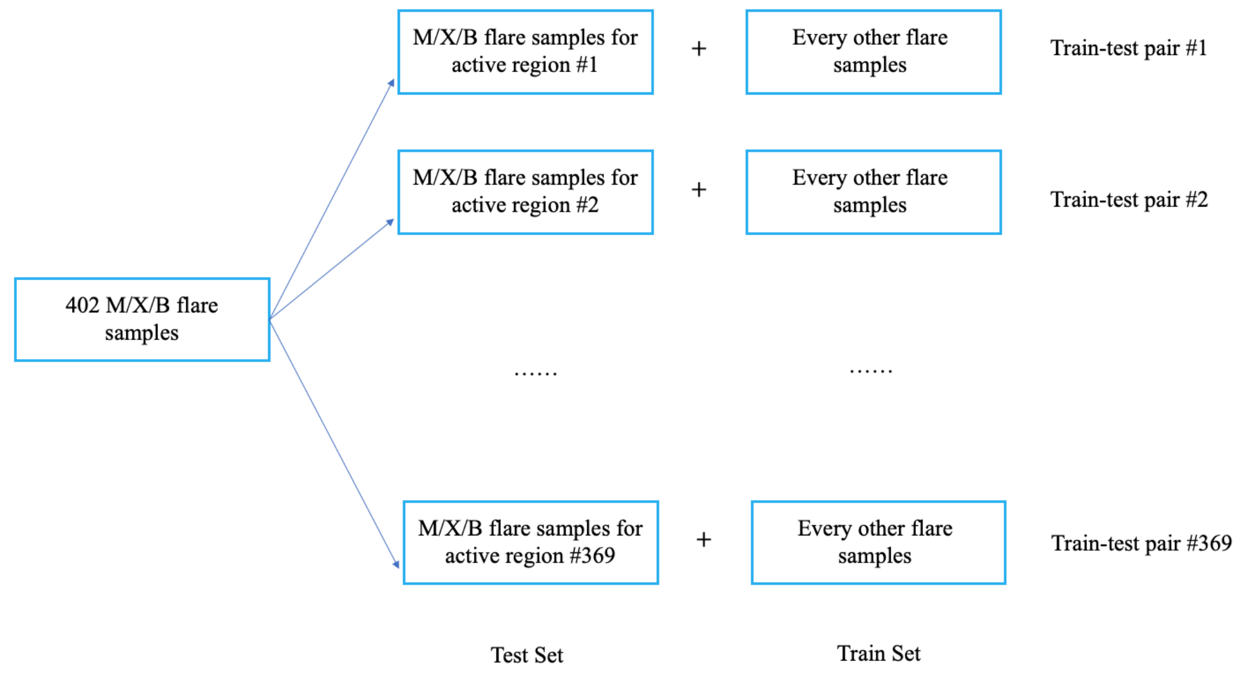
- Using the PIL data of Wang et al. (2019), which provides 20 physical parameters summarizing the local energy density, magnetic helicity, flux emergence and many other features of the local magnetic fields.

- Data: (x_i, y_i), x_i is 20*5 SHARP parameters along the PIL, $y_i = 1$ for M/X flares, $y_i = 0$ for B flares.
- The deep learning network we use is the Long-Short-Term-Memory (LSTM) model:


LSTM Network Architecture

Notation:
- $h_i^{<1>}$, $h_i^{<2>}$: LSTM output vector
- $c_i^{<1>}$, $c_i^{<2>}$: LSTM cell state
- X_i : 20*1 input vector
- Q : Prediction probability

- We train 369 LSTM models parallelly, with each model trained on a leave-one-out train set:


402 M/X/B flare samples

M/X/B flare samples for active region #1 + Every other flare samples — Train-test pair #1

M/X/B flare samples for active region #2 + Every other flare samples — Train-test pair #2

....

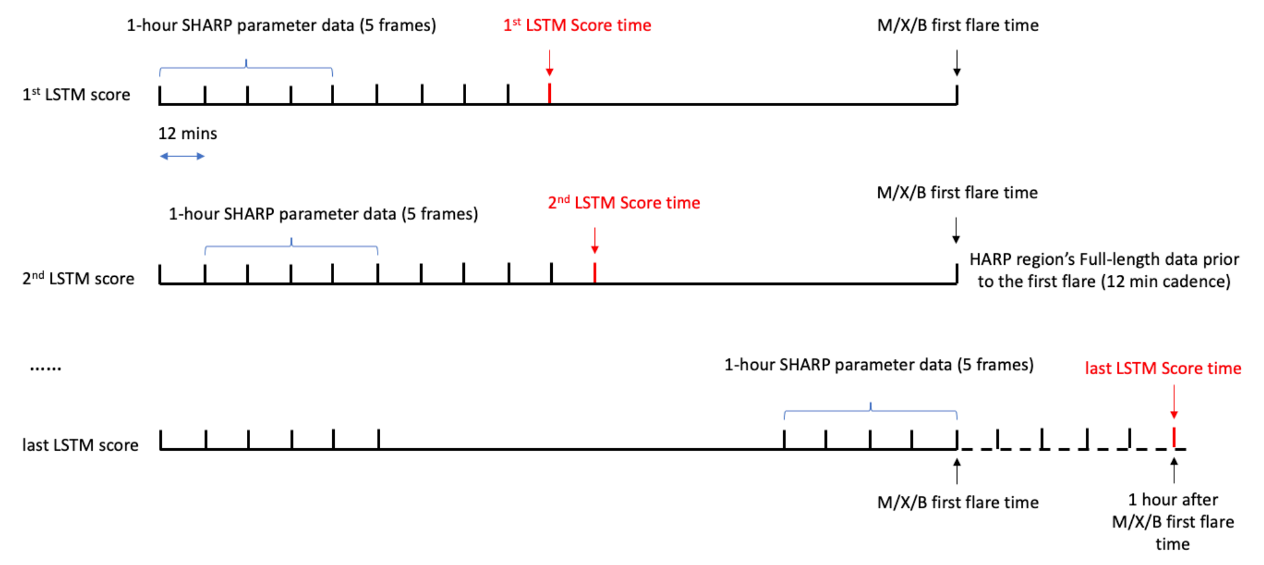
M/X/B flare samples for active region #369 + Every other flare samples — Train-test pair #369

Test Set

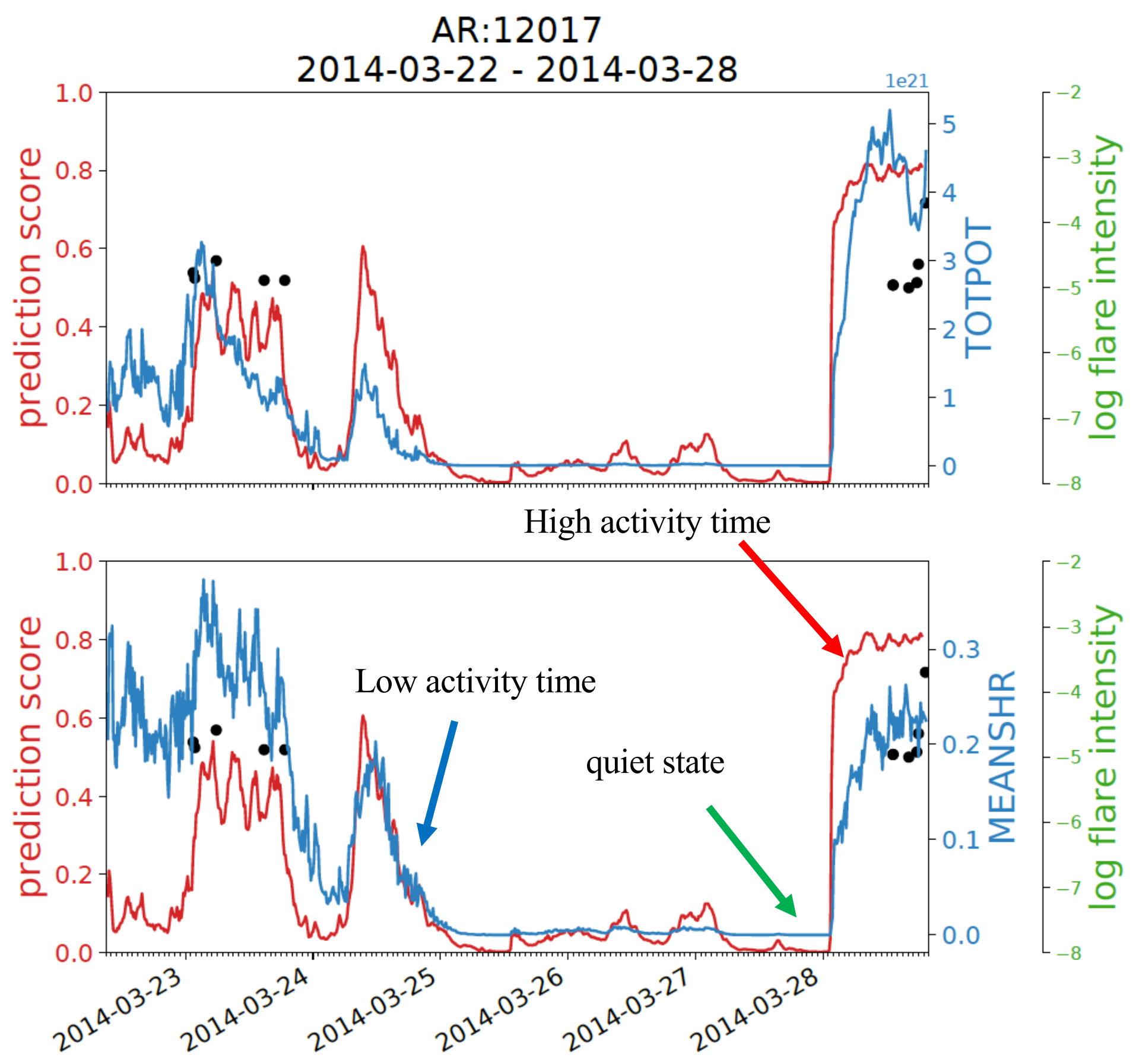
Train Set

LSTM Model Results

- For each flare in the test set, we generate a leave-one-out prediction curve with sliding window approach:


1-hour SHARP parameter data (5 frames)
1st LSTM Score time
M/X/B first flare time
1st LSTM score
12 mins
1-hour SHARP parameter data (5 frames)
2nd LSTM Score time
M/X/B first flare time
2nd LSTM score
HARP region's Full-length data prior to the first flare (12 min cadence)
......
1-hour SHARP parameter data (5 frames)
last LSTM Score time
last LSTM score
M/X/B first flare time
1 hour after M/X/B first flare time

- A case study on active region 12017:



- During 2014-03-28 1:00 AM, we see an abrupt score transition. Such a sudden transition of scores can be found preceding 35 M/X flares, averaging 48 hours before the flares.

- **Key question: what is the physical process underneath? Which features contain the signal that can stimulate the LSTM prediction score changes?**

LSTM Model Interpretation

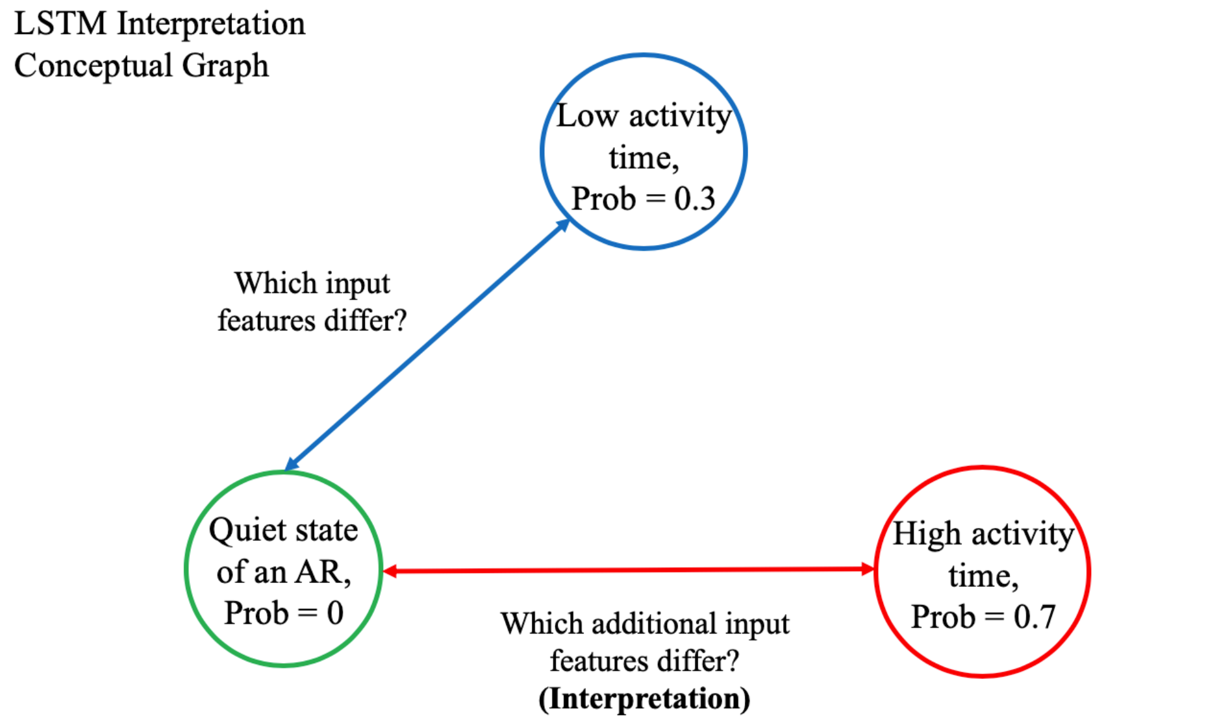
- To understand which physical process drives the machine learning prediction, we first focus on finding a "quite state" before each flare. **We choose the "quiet state" to be the time with the lowest LSTM score preceding the flare**(green arrow in the plot above).

- We obtain interpretations of LSTM model by comparing any LSTM inputs against the quite state input.

- Since LSTM input data are multi-dimensional time series, we define the following distance metric between any pair of LSTM inputs:

$$d(u,v) = [dtw(u_1, v_1), \dots, dtw(u_i, vi), \dots, dtw(u_{20}, v_{20})]$$

where u_i, v_i are the feature i of LSTM input u, v, and $dtw(.,.)$ is the Dynamic Time Warping (DTW) distance metric of two time series.


LSTM Interpretation Conceptual Graph

Low activity time, Prob = 0.3

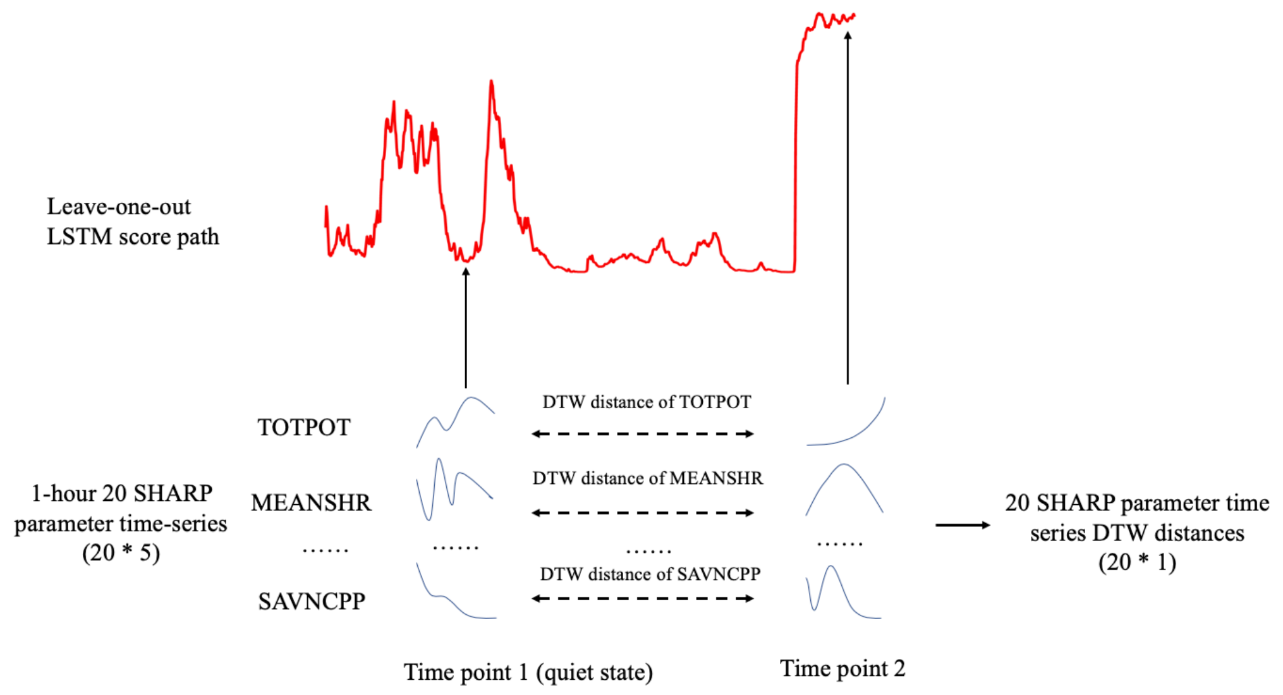
Which input features differ?

Quiet state of an AR, Prob = 0

High activity time, Prob = 0.7

Which additional input features differ? **(Interpretation)**

- We fix u to be the quiet state and only vary v:


Leave-one-out LSTM score path

1-hour 20 SHARP parameter time-series (20 * 5)

TOTPOT — DTW distance of TOTPOT

MEANSHR — DTW distance of MEANSHR

....

SAVNCPP — DTW distance of SAVNCPP

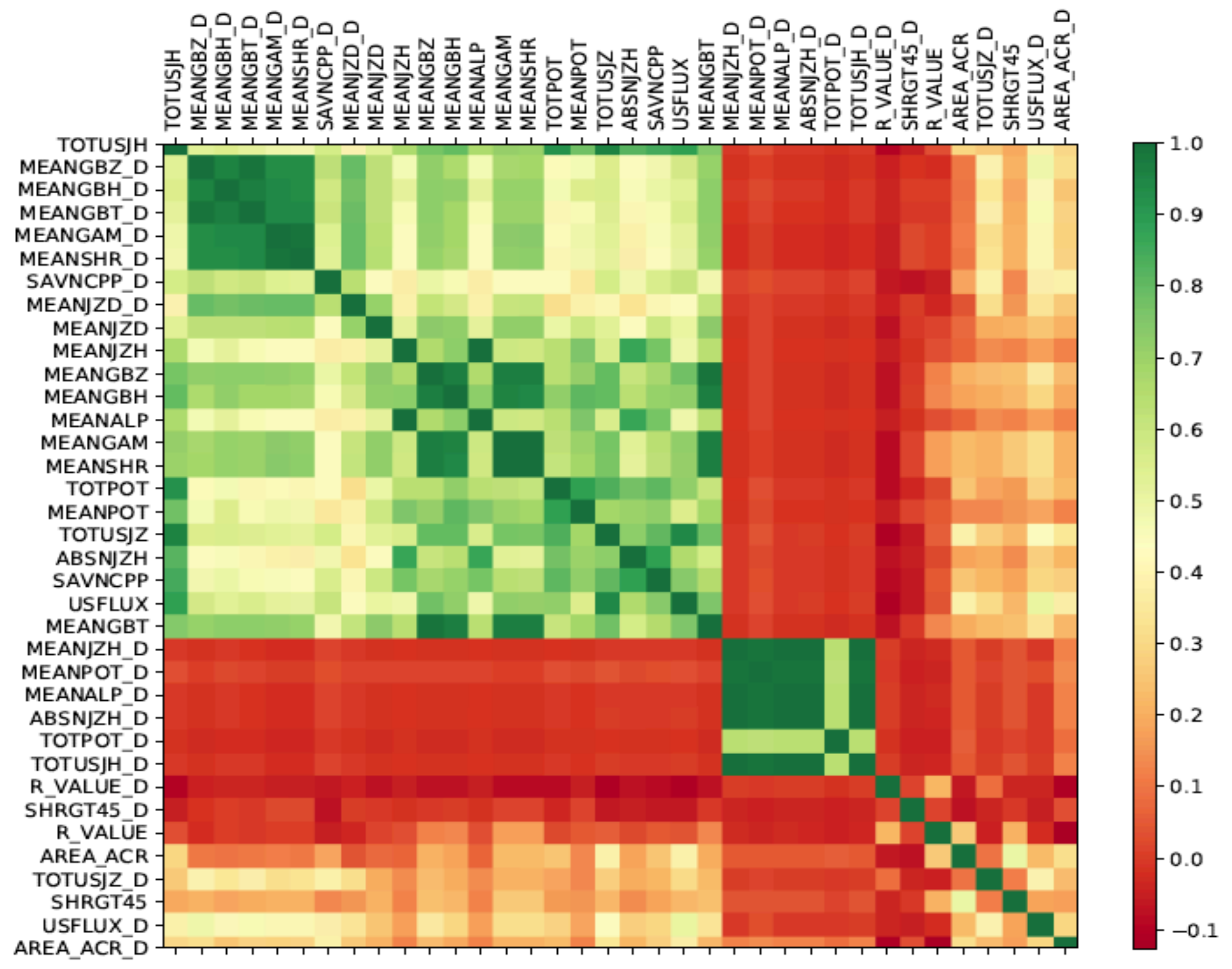
→ 20 SHARP parameter time series DTW distances (20 * 1)

Time point 1 (quiet state)

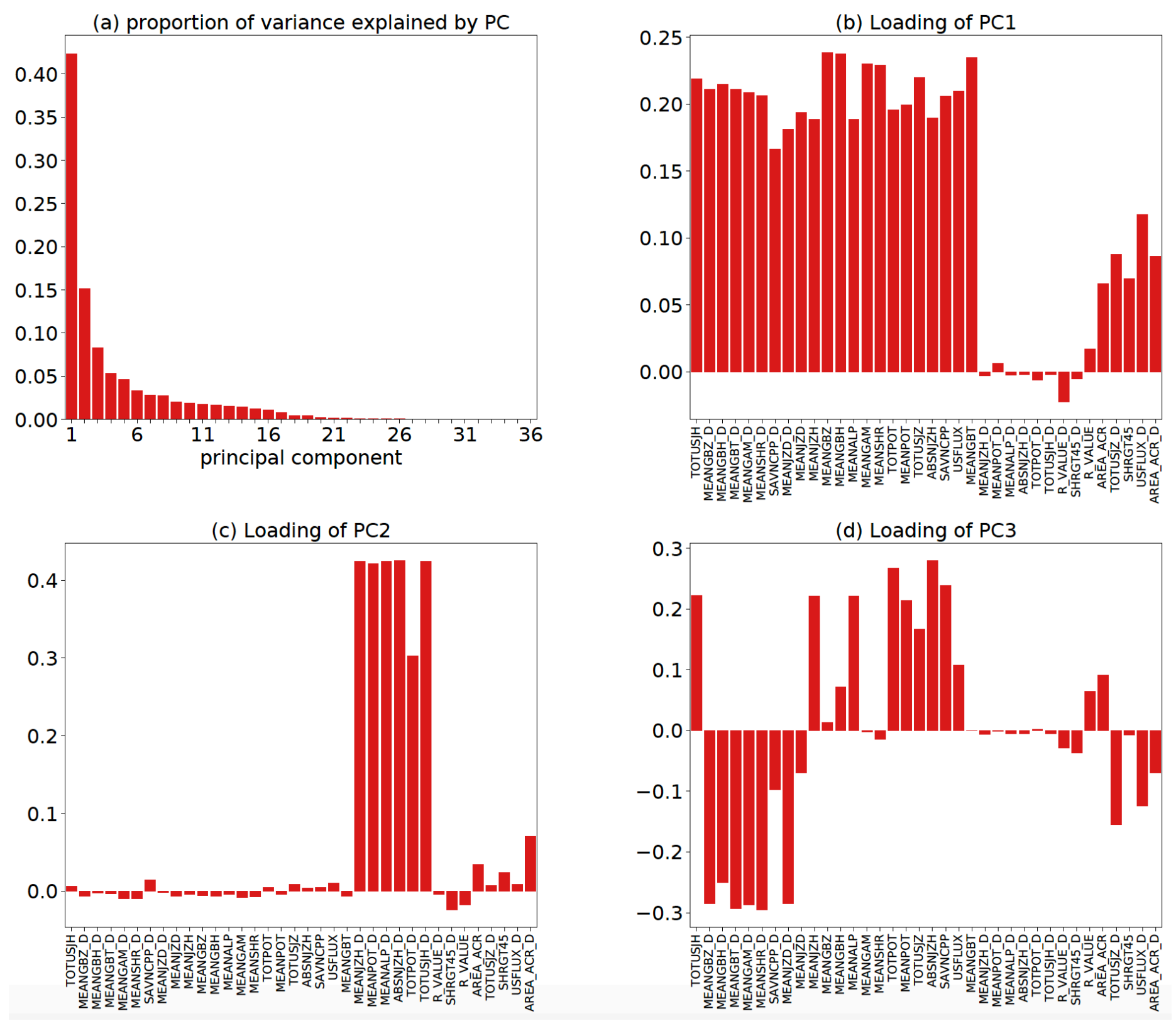
Time point 2

- We discard all flares whose "quiet state" has LSTM score above 0.2 since that is not considered as "quiet", and have 360 flare samples left.

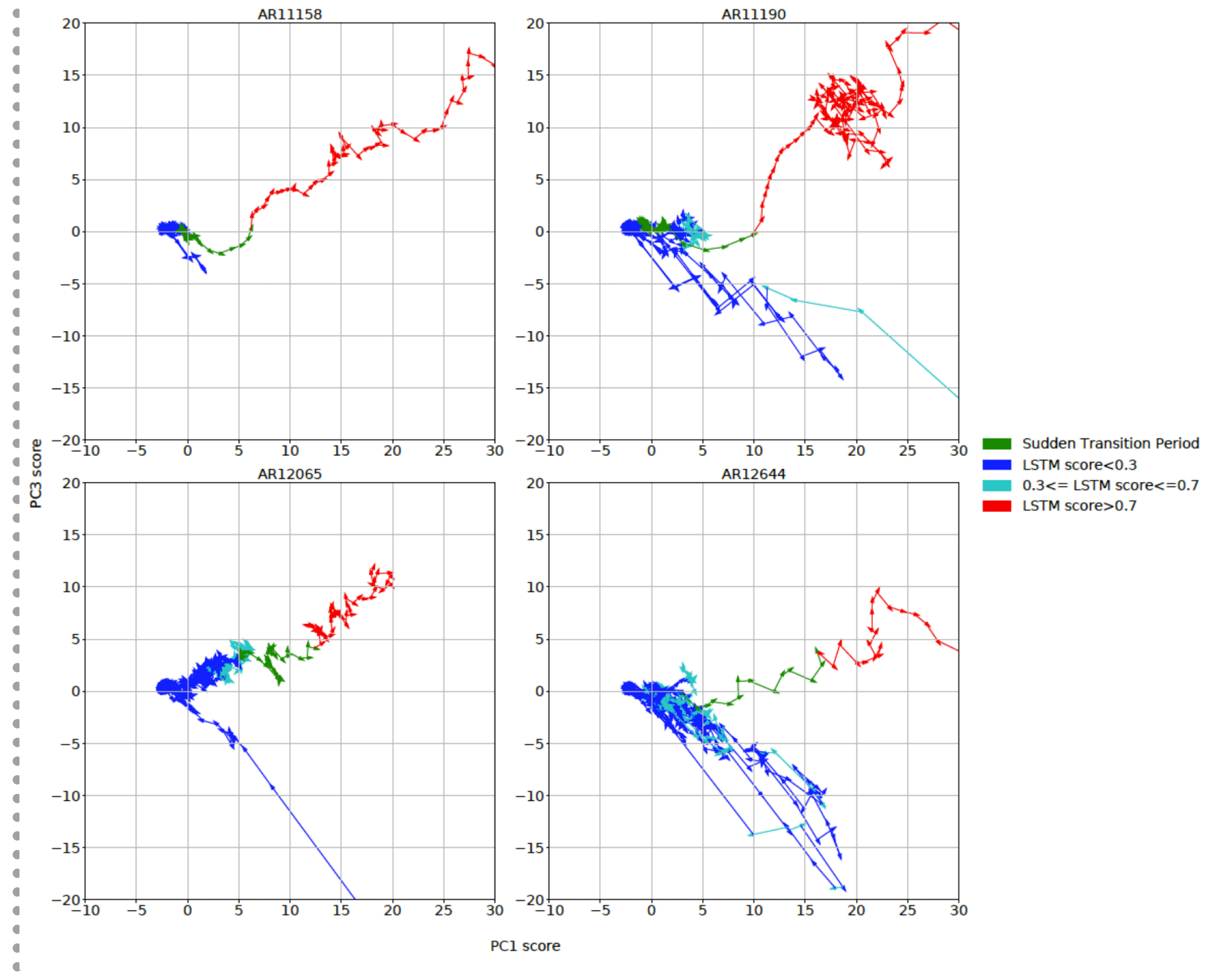
- To expand feature space, we also calculated the time derivatives of all 20 input features and computed the DTW distances for derivatives as well, the DTW distance features turn out to be highly collinear:



- We used Principal Component Analysis (PCA) to reduce the dimensions of all DTW distance feature:



Highly uniform patterns were found in the PC1-PC3 space for the M/X flares of many active regions:



Conclusion & References

- We propose a dimension-reduction technique based on DTW and PCA to summarize the information contained in matrix-shaped LSTM inputs. The low-dimensional representation of LSTM inputs shows some very interpretable learning patterns of LSTM model.

- SHARP parameters highly correlated with total free energy density are the important signals for strong flare eruption learnt by the LSTM model.

[1] Wang, J., Liu, S., Ao, X., Zhang, Y., Wang, T. and Liu, Y., 2019. Parameters Derived from the SDO/HMI Vector Magnetic Field Data: Potential to Improve Machine-learning-based Solar Flare Prediction Models. The Astrophysical Journal, 884(2), p.175.